

DOCUMENT RESUME

ED 311 867

IR 013 986

AUTHOR Shavelson, Richard J.; And Others
TITLE Evaluating Student Outcomes from Telecourse Instruction. A Feasibility Study.
INSTITUTION Rand Corp., Santa Monica, Calif.
SPONS AGENCY Corporation for Public Broadcasting, Washington, D.C.
REPORT NO ISBN-0-8330-0746-7; RAND/R-3422-CPB
PUB DATE May 86
CONTRACT 1813/25105
NOTE 80p.
AVAILABLE FROM The Rand Corporation, Publications Department, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138 (\$7.50).
PUB TYPE Reports - Evaluative/Feasibility (142)
EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Course Evaluation; *Distance Education; *Educational Television; *Evaluation Methods; Evaluation Problems; Feasibility Studies; Higher Education; Instructional Development; *Instructional Effectiveness; *Telecourses
IDENTIFIERS *Annenberg CPB Project

ABSTRACT

In November 1984, the Rand Corporation was asked by the Annenberg/Corporation for Public Broadcasting (A/CPB) Project to conduct a study during the spring 1985 school semester on the feasibility of evaluating student outcomes from A/CPB telecourses. This report documents the findings of that study, which had three goals: (1) to examine alternative designs for evaluating the impact of A/CPB telecourses on student outcomes; (2) to provide preliminary data on the impact of "The Constitution: That Delicate Balance" and "The New Literacy" on student outcomes; and (3) to recommend whether, within time and methodological constraints, it was feasible to evaluate student outcomes attributable to the A/CPB courses. Findings indicate that despite many difficulties, it is possible to monitor implementation so as to secure reliable data on student outcomes without intruding unduly into normal teaching routines. Certain conditions, however, must be observed. They include: (1) the use of a sufficient number of sites in order to oversample each course; (2) the provision of extensive lead time prior to implementation of the evaluation to allow site administrators and instructors to become involved in course design and instrumentation; (3) the mandating of pretesting for all students; (4) the collection of specific data on implementation at each site, including how the course was taught, the extent of integration of the video lessons into normal teaching, and political factors affecting course implementation; and (5) the provision of incentives to instructors and students to encourage them to adhere to evaluation procedures. (9 references) (GL)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

☐ Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

Evaluating Student Outcomes from Telecourse Instruction

A Feasibility Study

Richard J. Shavelson, Cathleen Stasz,
Steven Schlossman, Noreen Webb,
John Y. Hotta, Sandra Goldstein

Rand

"PERMISSION TO REPRODUCE THIS
MATERIAL IN MICROFICHE ONLY
HAS BEEN GRANTED BY

Bill Bower

The research described in this report was sponsored by the Corporation for Public Broadcasting under Contract No. 1813/25105.

ISBN: 0-8330-0746-7

The Rand Publication Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

R-3422-CPB

Evaluating Student Outcomes from Telecourse Instruction

A Feasibility Study

Richard J. Shavelson, Cathleen Stasz,
Steven Schlossman, Noreen Webb,
John Y. Hotta, Sandra Goldstein

May 1986

Prepared for the
Corporation for Public Broadcasting



PREFACE

American educational television has had a checkered history. First conceived mainly as a vehicle for transmitting classroom lectures or demonstrations to large numbers of students or to home audiences, it took as its basic format the "talking head." Today's producers of educational television—from preschool to postgraduate levels—use sophisticated production technologies creatively to transcend traditional teaching strategies and to capitalize on the distinctive attributes of the medium. Newer telecourses must satisfy the high expectations of audiences with nearly two generations of television viewing experience.

The Annenberg/Corporation for Public Broadcasting (A/CPB) Project recently created a series of telecourses covering such topics as the inner workings of the human brain and of the Congress, the rudiments of physics, and the uses and misuses of the computer. In late November 1984, The Rand Corporation was asked by A/CPB to conduct a study during the spring 1985 school semester on the feasibility of evaluating student outcomes from A/CPB telecourses. This report documents the findings of a study funded by A/CPB that had three goals:

- To examine alternative designs for evaluating the impact of A/CPB telecourses on student outcomes
- To provide preliminary data on the impact of *The Constitution: That Delicate Balance* and *The New Literacy* on student outcomes
- To recommend whether, within time and methodological constraints, it is feasible to evaluate student outcomes attributable to the A/CPB courses

The findings should be of interest to persons engaged in developing, teaching, implementing, and evaluating future telecourses.

SUMMARY

Telecourses provide instruction primarily to diverse populations of students off-campus, most notably through television, but via other media as well. Although the term *telecourse* implies that television is the predominant instructional medium in the course, this is not usually the case. Telecourses are typically highly diversified, multimedia educational packages in which television may play a variable pedagogical role. In addition, telecourses vary as widely as do courses using only traditional media on dimensions other than television use. As a result, the "same" telecourse may have widely varying versions depending on features of its implementation at different sites or by different instructors.

The educational value of telecourses remains in dispute. Research has been unable to show that telecourses offer any educational advantages over traditional media for the same populations of students. Moreover, studies have not adequately addressed the "exchangeability" of student outcomes in traditional courses and telecourses: in other words, it has not been established that the knowledge, skills, and attitudes acquired by students in a telecourse are comparable to ("exchangeable for") the knowledge, skills, and attitudes acquired by students in an equivalent traditional course.

This study addressed three questions about evaluating student outcomes from telecourse instruction: (1) Which designs are best suited to evaluate student outcomes? (2) To what extent do two telecourses—*The Constitution: That Delicate Balance* and *The New Literacy*—result in students acquiring new knowledge and skills and is this learning exchangeable for what students in traditional courses learn? And (3), given time and methodological constraints, is it feasible to evaluate student outcomes in telecourses?

ALTERNATIVE DESIGNS FOR EXAMINING STUDENT OUTCOMES FROM TELECOURSES

The evaluator seeking to examine student outcomes from telecourses has a wide range of alternative designs from which to choose. Which choice is best for a given situation depends on many factors, not the least of which are the types of decisions (and decisionmakers) on which the evaluation focuses and the feasibility of implementing the design.

Audiences for telecourse evaluations include instructors/academic departments, telecourse administrators, funding agency administrators, and telecourse developers. All consider two questions: (1) How much do students gain in knowledge, skills, and attitudes from the telecourse? (2) Is what they learn the same as what students learn in traditional courses that are intended to teach the same subject matter? The audiences differ, however, in the relative weight they give these questions. Because telecourse developers seek information on improving the course, for example, they probably weight the question of knowledge gain as more important than the question of equivalence of outcomes. Academic departments/instructors, on the other hand, are ultimately concerned with academic standards, and are thus concerned that the telecourse produce outcomes equivalent to traditional courses.

While we acknowledge the importance of assessing student gains, we believe that telecourse evaluations must ultimately focus on outcomes and address the *exchangeability* of these outcomes with those attained by students in traditional courses. By "exchangeability" we mean the extent to which the knowledge, skills, and attitudes acquired by students from a telecourse are interchangeable with the knowledge, skills, and attitudes that are: (a) valued by faculty and administrators, and (b) acquired by students enrolled in the same course offered as part of the traditional curriculum. The concept of exchangeability reflects the central question raised by policymakers: "Do students learn from the telecourse what students in a traditional course learn?" Exchangeability also introduces a new perspective on the old axiom derived from past media comparison research: "media don't make a difference" in achievement. If media "don't make a difference," there is good reason to believe that the answer to policymakers' exchangeability question is, "yes, alternative courses are exchangeable." But past research has only addressed the question of exchangeability for students drawn from the *same* population (either by random assignment or (inadequate) statistical adjustment). For populations nested in tele- and traditional courses, the answer might differ. Finally, exchangeability translates directly into a set of questions that guide evaluation: (a) Does the telecourse content meet academic standards and, if so, do the students acquire the knowledge and skills imparted? (b) Do students in a telecourse learn as much as do students in traditional courses that cover the same objectives, assignments, and materials? and (c) Do students in a telecourse learn as much as students in another course that fulfills the same curricular goals but does not necessarily use the same assignments and materials?

The criterion of exchangeability influences the selection of an evaluation design: (1) The design should allow for student choice of

course type because students choosing telecourses differ from those choosing traditional courses, otherwise, attrition may be nonrandom; (2) the design should yield an estimate of student gain from telecourses; and (3) the design should provide information about the *extent* of students' knowledge, skills, and attitudes at the end of the course, compared with those of students in traditional courses.

Implementation feasibility also influences choice of design. Since institutions of higher education have extensive academic freedom, the greater the control over the educational setting required by the design, the more difficult it is to implement. In addition, adherence to planned treatments as embodied in tele- and traditional courses and instrumentation will be difficult to protect, especially if the evaluation team needs to place constraints on both.

Thus, on both feasibility and student-choice grounds, random assignment of students and/or instructors to tele- or traditional courses will probably (a) change the nature of the treatments, (b) change the nature of the student population, or (c) both. In addition, the design *must* include an implementation study, in which the researcher is actively involved in the evaluation on-site.

DESIGN RECOMMENDATIONS

Once randomized designs have been ruled out, three designs can be identified as relevant to evaluating student outcomes from telecourses. In Design 1, *Uncontrolled Assignment to Form Non-Equivalent Groups*, students self-select into tele- or traditional courses. Before and after treatments, knowledge, skills, and attitudes are measured. This design compares intact classes and recognizes that the student population cannot be separated from course implementation.

Design 2, the *Patched-Up Design*, is appropriate when institutions regularly cycle students through the same course, such that students from one cycle can serve as a control group for students from another cycle. Institutions must offer the same telecourse in consecutive semesters of the same year; achievement and, if possible, other data must significantly overlap across the two semesters. Comparisons of fall posttest data (treatment group) with spring pretest data (control group), for example, help determine whether students learned more from taking the telecourse than they would have learned without it. Spring pretest-to-posttest comparisons are made to replicate the treatment effect and to gain confidence that some unidentified factor did not cause this change in performance.

The third design, *Case Study Methods*, provides narrative, usually descriptive, accounts of an object of social inquiry, such as a classroom, a school system, or other bounded system and focuses on the cultural context of that system. Case study methods are important for understanding tele- and traditional course outcomes, but they are unlikely to resolve questions about *exchangeability*; they should be used in conjunction with other designs.

IMPLEMENTATION FINDINGS

Implementation for the two telecourses—*The Constitution* and *The New Literacy*—was highly variable and demonstrates the necessity of embedding analysis of student outcomes in a richly textured description of the implementation process. Variability characterizes the way the courses were implemented, the characteristics of the students served and their course participation, and the problems and difficulties encountered by the evaluators. These results support the notion that each telecourse "treatment" is different and is defined by the telecourse and its local implementation.

We gathered data at three sites for *The Constitution* and at five sites for *The New Literacy*, including pre- and posttests, background and postcourse questionnaires, and information about course implementation (through interviews and on-site observations).

THE CONSTITUTION

As designed and distributed, *The Constitution: That Delicate Balance* is a one-semester introduction to constitutional law and public policy. Course materials consist of 13 one-hour videolessons, a textbook, and a study guide.

Variability in **course implementation** was demonstrated by differences in cost (students at one site paid approximately ten times as much as at other sites); format (all viewed the telecourse video components, but students at one site read only the study guide); requirements (one site gave four examinations instead of two); and instructors (both experienced and inexperienced in teaching telecourses).

Two important problems influenced **evaluation implementation**: (1) a small sample size and (2) the intrusive nature of evaluation instruments with objective test items in a course where instructors ordinarily prefer short answer and essay exams. Although there were few administrative hitches at these sites, lack of communication and other problems caused the instructor at one site to delay turning in test

data. At another site, to the surprise of the researchers and the administrator, the instructor administered our examination *after* his own three-hour, six-essay final examination. As a result, student test data were compromised.

Data on **student characteristics** from two sites indicate that telecourse populations are not homogeneous. Students at Site B tended to be younger, work less, and have higher incomes and educational attainment than students at Site A. Site B had a larger percentage of minority students; they had lower verbal aptitude and attitudes toward telecourses, less telecourse experience, and were more likely to drop out.

Differences between sites in **telecourse participation** were also evident. Site B students reported, on average, attending fewer meetings, getting less out of those meetings, and studying less than students at Sites A or C. Because of instructor differences, Site B students did not take notes on the video portion of the course. Fewer Site B students discussed the program following its viewing, although they were more likely to contact their instructor for assistance than students at the other sites.

THE NEW LITERACY

The New Literacy is a one-semester introduction to computers. It consists of 26 half-hour videolessons that were created out of the textbook, and is accompanied by a study guide to integrate the videolessons with the text. At four of five study sites, the textbook took precedence over the videolessons for determining course content.

As with *The Constitution*, variability was evident in **course implementation**. Site D was of special interest because it offered in addition to *The New Literacy* telecourse (D1), a traditional introductory computer course (D2) equivalent in objectives, credits, and required text. All telecourse sites required orientation and met about five times, except for Site H, which met as frequently as the traditional course at Site D2 (18 meetings). In addition to the midterm and final examinations required at all telecourse sites, Site D1's instructor required students to complete and return six quizzes and gave students the opportunity to receive extra credit by completing miniprojects or attending a workshop. Site H's students took six in-class quizzes and field trips. Instructors varied from openly disgruntled to highly professional, and telecourse administration from highly supportive to suspicious of outside evaluators.

By far, the most serious problem for **evaluation implementation** occurred at Site E, where the instructor conducted the midterm and final review by reading each examination question in order and providing the correct answers. Not surprisingly, these students did fantastically well on the examinations! Examinations developed by the research team were the chief cause of problems at Site D and resulted in instructor dissatisfaction and changes in course format.

Data on **student characteristics** show that student populations served by *The New Literacy* systematically differ in age, percentage of women in the course, percentage of students holding bachelor's degrees, and verbal ability. Sites are similar in attitudes, reported grades, enrollment in degree programs, enrollment in the course for credit, and computer background and interest.

Variation in course implementation produced variation in **course participation**. Students at Site H, which held significantly more meetings than the other telecourses, rated the value of these meetings higher and met outside of class more often than students at other sites. "Studenting" at the other four sites is quite similar, with unsystematic exceptions. Students at all sites gave moderately high ratings to overall preparation of the instructor, the worth of the telecourse, and the textbook and study guide.

In sum, no two sites were even close to being alike. Courses differed in format, requirement, grading schemes, instruction, and administration. In fact, implementation varied so much that only one evaluation approach proved feasible.

STUDENT OUTCOMES

We could not evaluate exchangeability of tele- and traditional course outcomes because of evaluation implementation problems and because very few appropriate comparisons could be found for these two telecourses. As a result, we were left with the patched-up design in *New Literacy* sites as the only design that could provide information about student knowledge gain. To implement this design at four *New Literacy* sites (E, F, G, and H), we developed a common set of items that, together with a subset of locally developed items, constituted the pretest for each site. Common items also appeared as a part of each site's regular midterm and final examinations. Except for Site E (where the instructor gave test answers during review sessions), comparisons of the fall semester posttest scores (treatment group) with the spring semester pretest scores (control group) indicate a consistent difference of about 8 points out of a total possible 50 at each site.

Again, except for Site E, a similar gain is found from spring pretest to posttest at each site. As implemented at three sites, *The New Literacy* telecourse significantly increases students' knowledge about computers.

To compare the consistency of these effects between sites, we converted the mean difference score at each site into a measure of effect size. A consistent effect size of about 1.36 standard deviations comparing fall posttest and spring pretest scores is evident across all sites. Except for Site E, a similar effect is found for the spring pretest-to-posttest comparison. Effect size for the differences between fall posttest and spring posttest scores, as expected, is negligible across sites (except for E). Thus, it appears that gains in test scores are remarkably stable from one semester to another, despite variation in telecourse implementation within and across sites and different examinations. We obtain similar results using only the common items (an identical measure for all sites) to support our conclusion that students learn from the telecourse.

Correlations between student course participation ("studenting") variables and common items show three studenting variables consistently related to test performance across five sites: (a) percentage of textbook reading assignments completed, (b) number of class meetings attended, and (c) rating of instructor's preparation. Except for Site H, student attitudes toward telecourses declined slightly from pretest to posttest, although they held, on average, positive attitudes.

CONCLUSIONS AND RECOMMENDATIONS

Study results lead us to conclude that, methodologically, the patched-up design provides a relatively unobtrusive and time-minimizing design that produces reliable, valid information for evaluating student outcomes from telecourses. This design, as implemented, cannot determine whether what students learn from telecourses is exchangeable for what is learned in an equivalent traditional course. Nevertheless, it still provides important information to various decisionmakers.

Despite abundant difficulties, we believe it possible to monitor implementation in a manner that secures reliable data on student outcomes without intruding unduly into normal pedagogical routines. Certain conditions, however, must be observed.

1. A sufficient number of sites must be found to oversample each course.
2. Extensive lead time must precede site selection.

3. Pretesting must be mandated and carried out for all students.
4. Implementation data, integral to analysis of student outcomes, should be gathered by on-site researchers. Four key questions are:
 - a. How was the course actually taught?
 - b. Did the instructor integrate the videolessons into his teaching and evaluation?
 - c. How were the review sessions conducted?
 - d. What political factors affected course implementation?
5. Via education and, perhaps, monetary inducements, instructor and student willingness to adhere to all evaluation procedures must be secured.

Although rigorous evaluations of telecourses can be hampered by serious impediments, the impediments are not insurmountable.

ACKNOWLEDGMENTS

Throughout this research we benefited from the advice and counsel of Peter Dirr of the Corporation for Public Broadcasting. Sally Beatty of the Southern California Consortium for Community College Television provided invaluable help in identifying potential sites and in other aspects of the project. Professor William Geeslin gave valuable research assistance at a distant site.

Steven Klein, a Rand colleague, and Richard Clark, of the University of Southern California, provided insightful, critical reviews of earlier versions of this report. The authors implemented their suggestions with the help of Rick Eden, a member of Rand's Communications Consulting Group. Jan Iverson, Linda Daly, Jean Thomas, and Alyce Raphael provided valuable assistance in support of the study.

We are especially grateful to the school administrators, instructors, and students who gave generously of their time and allowed us to conduct our evaluation on their campuses. Without their assistance, this study would not have been possible. Although we depended on such diverse and willing assistance, we alone are responsible for any shortcomings.

CONTENTS

PREFACE	iii
SUMMARY	v
ACKNOWLEDGMENTS	xiii
TABLES	xvii
Section	
I. INTRODUCTION	1
II. ALTERNATIVES FOR EVALUATING STUDENT OUTCOMES FROM TELECOURSES	5
Audiences for Telecourse Evaluations	5
Exchangeability: A Unifying Concept for Evaluating Student Outcomes	7
Implementation Feasibility	9
Design Recommendations	10
Inappropriate Designs	11
III. A TALE OF TWO TELECOURSES	14
Data Gathering Procedures and Instruments	14
Overview of Implementation	15
<i>The Constitution</i>	15
<i>The New Literacy</i>	22
Summary	30
IV. STUDENT OUTCOMES FROM TELECOURSES	33
Patched-Up Design—The Recurrent Institutional Cycle Design	34
Student Outcomes	36
Concluding Comments	46
V. RECOMMENDATIONS AND CAVEATS	47
1. Oversample ... to Excess	48
2. Obtain Ample Lead Time	49
3. Pretest Comprehensively	50
4. Study Course Implementation in Depth	52
5. Do Not Underestimate the Intrusiveness of Evaluation Research	55
Closing Observations	57

Appendix: RAND COMMON ITEMS	59
REFERENCES	63

TABLES

1. Telecourse implementation at <i>The Constitution</i> sites	17
2. Evaluation implementation problems at <i>The Constitution</i> sites	17
3. Course participation at <i>The Constitution</i> sites	21
4. Telecourse implementation at <i>The New Literacy</i> sites	23
5. Evaluation implementation problems at <i>The New Literacy</i> sites	26
6. Student characteristics at <i>The New Literacy</i> sites	29
7. Course participation at <i>The New Literacy</i> sites	31
8. Internal consistency of vocabulary and achievement tests . .	35
9. Comparison of item difficulty between unique and common items	43
10. Selected correlations between student characteristics and final test scores	43
11. Selected correlations between student course participation and final test scores	45

I. INTRODUCTION

A telecourse provides instruction primarily to diverse populations of students off-campus, most notably through television, but via other media as well. Although the term *telecourse* seems to imply that television is the predominant instructional medium in the course, this is not usually the case. Telecourses are typically highly diversified, multimedia educational packages in which television may play a variable pedagogical role. In addition, telecourses vary as widely as courses using only traditional media on dimensions other than television use. As a result, the "same" telecourse may have widely varying versions depending on features of its implementation at different sites or by different instructors.

The motivations for telecourses in American higher education—as a medium for college-level instruction—are many:

- Large numbers of people can be served.
- A wide variety of people can have access to higher education, perhaps not otherwise possible.
- The television medium admits to a wider variety of instructional "treatments" (multimedia communication of knowledge and attitudes) than does traditional instruction.
- The technical quality of the video components available for use in telecourses has improved markedly over the years, and the best educational television uses today's most sophisticated production techniques.

Nevertheless, the educational value of telecourses remains in dispute. Research has been unable to show that telecourses offer any educational advantages over traditional media for the same populations of students. Moreover, studies have not adequately addressed the "exchangeability" of student outcomes in traditional courses and telecourses: in other words, it has not been established that the knowledge, skills, and attitudes acquired by students in a telecourse are comparable to ("exchangeable for") the knowledge, skills, and attitudes acquired by students in an equivalent traditional course. Finally, the production costs of video components are high compared with those of components in other media, and the number of students served is often smaller than planned.

The telecourse is a controversial, emotionally charged issue in higher education. To some it represents a threat—indeed, the greater the

sophistication of the course, the greater the competition and threat to traditional educational institutions, their curricula, and instructors. American higher education institutions have traditionally prided themselves on their local and regional character and responsiveness. Telecourses, by contrast, represent a nationalizing, translocal, and transregional influence. Hence, the relation between telecourse producers and educators is inevitably tension-ridden. Even if telecourses do carry their own weight financially on campus, they may challenge both the local initiative ethos and survival needs of America's higher education industry.

Educational policymakers attempting to decide whether or how to incorporate telecourses must weigh many considerations, including cost and educational efficacy. What kinds of students do telecourses attract? Are they comparable in ability and in what they learn in telecourses to the students attracted to traditional courses? Are they more or less likely to drop out? Do telecourses draw students away from traditional courses? Are instructors threatened by the new role imposed on them by the television medium or by the possibility of losing their jobs? How much will the video components cost to use? Will nonstudents in the community benefit from the broadcast of the video components? And, most importantly perhaps, are the courses comparable in educational value to similar courses using traditional media? That is, do students enrolled in telecourses learn what traditional college students learn from traditional courses?

This study attempts to address a subset of these concerns, those dealing with the educational outcomes of telecourses. It examines the feasibility of evaluating the impact of telecourses on students' acquisition of knowledge, skills, and attitudes toward the subject-matter of the telecourse. More specifically, the study had three goals:

- Examine alternative evaluation designs to assess student outcomes;
- Provide preliminary data on student outcomes for two telecourses: *The Constitution: That Delicate Balance* and *The New Literacy* (an introduction to computers); and
- Recommend whether it is feasible to evaluate student outcomes within time, institutional, and methodological constraints.

If reliable, valid evaluations of student outcomes in telecourses are feasible and affordable, they could provide information useful both to educational policymakers who ask whether the telecourse they are paying for is at least as "good" as the traditional course; to telecourse designers who ask whether students gain as much as possible in

knowledge, skills, and attitudes from their telecourse; and to administrators and faculty who ask both questions.

This study addresses evaluation of outcomes only for students enrolled in colleges and universities for telecourse credit. The number of such students is far smaller than the number viewing them. (This is especially true for *The Constitution*, which was televised nationally in fall 1984.) The students enrolled in the telecourse at a higher education institution presumably represent only the tip of the iceberg. This study is mute on the feasibility of evaluating outcomes either for "students" who do not enroll but do view the telecourse and read the textbook at home, or, by far the larger number, those who do not enroll but only view the telecourse (perhaps occasionally). This larger audience might very well be the focus of a subsequent feasibility study.

In this study, we examined student outcomes at three sites for *The Constitution* and at five sites for *The New Literacy*, with a number of evaluation designs. We gathered data on telecourse implementation (e.g., course requirements, instructor experience, administrative policies), student characteristics (including previous experience, ability, and demographics), course participation, attitudes, and learning. In the end, we could only determine student outcomes (attitudes, participation, and learning) from *The New Literacy*. We conclude that this telecourse, as implemented at four sites, significantly increases students' knowledge about computers.

Our findings demonstrate the extraordinary diversity of telecourse implementation. No two sites were even close to being alike. Course formats, requirements, and grading schemes varied widely in design and even more in operation; instructors differed dramatically from enthusiastic to disgruntled, from highly professional to egregiously unprofessional. Pat "lessons" from these myriad implementation experiences are clearly unjustifiable. Indeed, if one grand "lesson" had to be drawn, it would be to expect extreme variation and unreliability in operationalizing research designs in the American "system" of higher education.

The study found that the sites at which the two telecourses were offered varied so much in their implementation of the courses that only one evaluation approach proved feasible. This approach, known as the "patched-up" design, could be applied where (a) the same telecourse was offered during fall and spring semesters in the same year, and (b) overlapping achievement data are available for fall final examinations and spring pretest and final examinations. Comparing fall posttest with spring pretest achievement helps answer the question: "Do students learn more from a telecourse than they would have learned without it?" Comparing spring pretest with posttest achievement

provides an opportunity to replicate the "treatment effect" and increase confidence that the course itself rather than some unidentified factor caused the change in performance.

Serious impediments to implementing rigorous evaluations of telecourses are many, but they are not, in our judgment, insurmountable. Despite abundant difficulties, we believe it is possible to monitor implementation so as to secure reliable data on student outcomes without intruding unduly into normal pedagogical routines. Certain conditions, however, must be observed:

1. A sufficient number of sites must be available to oversample each course.
2. Extensive lead time must precede site selection.
3. Pretesting must be mandated and carried out for all students.
4. Implementation data must be integral to analyses of student characteristics. Researchers should be on-site to gather information on four key matters:
 - a. How was the course actually taught?
 - b. Did the instructor integrate the videolessons into his teaching and evaluation?
 - c. How were the review sessions conducted?
 - d. What political factors affected course implementation?
5. Through education and, perhaps, monetary inducements, instructor and student willingness to adhere to all evaluation procedures must be secured.

This report is divided into five sections. Section I is the Introduction. Section II identifies appropriate and inappropriate evaluation designs and explains the notion of "exchangeability," which can help guide evaluation studies. Section III describes the design and implementation of the two telecourses we studied. Section IV describes the findings of the patched-up design, the only evaluation approach feasible in light of the variability in telecourse implementation. Section V offers recommendations for the design of future telecourse evaluations.

II. ALTERNATIVES FOR EVALUATING STUDENT OUTCOMES FROM TELECOURSES

The evaluator seeking to examine student outcomes from telecourses has a wide range of alternative designs from which to choose (e.g., Campbell and Stanley, 1963; Cook and Campbell, 1979; Streuning and Guttentag, 1975). Which choice is best for a given situation depends on many factors. Two of the most important factors are: (a) the types of decisions (and decisionmakers) on which the evaluation focuses, and (b) the feasibility of implementing the design.

The first part of this section discusses the different types of decisions on which evaluation results bear, and presents the notion of exchangeability. The next part describes the various types of designs that can be used for each purpose, and the factors that influence the feasibility of implementing each design. Although we focused our feasibility study on telecourses offered by higher education institutions for credit, other studies should attend to the potentially much larger "student" audience not enrolled for credit. In addition, although we assume that the most important telecourse outcome is subject-matter achievement, other outcomes are important as well, including attrition, attitudes toward the course, and satisfaction with the course.

AUDIENCES FOR TELECOURSE EVALUATIONS

Audiences for telecourse evaluations include: (a) instructors/academic departments, (b) telecourse administrators, (c) telecourse developers, and (d) funding agency administrators. Regardless of audience, two questions are raised: How much do students gain in knowledge, skills, and attitudes from the telecourse? Is what they learn the same as what students learn in traditional courses that are intended to teach the same subject matter? (That these two questions are different can be seen from the fact that students may gain substantial knowledge in a course but still fall short of the knowledge required by academic standards.)

These audiences differ in the relative weight they give these questions. Telecourse developers probably weight the question of knowledge gain as much more important than the question of equivalence of outcomes. That is, they seek information on improving

the telecourse. Funding decisionmakers, we suspect, weight the importance of the two questions just the opposite. The bottom line is whether the "product" they are paying for is as "good" as the traditional course. Academic instructors/departments, in our experience, consider both questions to be important but ultimately hold the telecourse responsible for producing outcomes equivalent to traditional courses. They seek, in large part, to maintain academic standards. Finally, telecourse administrators hold both questions to be important, but also have to deal with the economic reality that telecourses often must support themselves. To be self-supporting, telecourses must address student populations that are often different from student populations traditionally served in undergraduate college programs. In most cases, nontraditional student populations have, on average, poorer academic preparation than traditional students and less time to study. For these administrators, then, the question of knowledge gain may be more important. Even if the telecourse outcomes are not equivalent to the traditional-course outcomes—and there is good reason to expect that, given student population differences, they will not be—a large number of people will, for the first time, have access to higher education. Even if only a small percentage of telecourse students perform to traditional course standards, this achievement, amortized over large telecourse enrollments, is sufficient for administrators to justify the telecourse option.

With all of these audiences, how is one to choose an evaluation design? We begin to narrow the range of alternatives by noting that curriculum developers have a fundamentally different agenda than the other three audiences, an agenda better addressed by collecting extensive information on which telecourse components do or do not "work" (i.e., formative evaluation). The other audiences' agendas all deal with a comparative question: "Given traditional courses and telecourses, should the traditional or both be used to provide higher education?" This study considers summative telecourse evaluation.

Agendas still vary among the remaining audiences. Should the evaluation design focus on gains or outcomes? We believe that gain in knowledge is a necessary condition for an evaluation of student outcomes in telecourses. Decisionmakers need to know that telecourses achieve significant gains in knowledge. But alone, information about gains cannot satisfy the important criterion of academic standards. To know that students increased their knowledge substantially does not tell the decisionmaker whether the students possess the knowledge, skills, and academic attitudes requisite for satisfactory completion of the course. For this reason, we believe that evaluations of telecourse outcomes must ultimately lie in student outcomes and address the

exchangeability of these outcomes with those attained by students in traditional courses. The outcomes of traditional courses, then, establish a criterion of academic standards against which telecourse outcomes should be compared.

On the basis of audience, we can identify several criteria for selecting an evaluation design:

1. The design should allow for student choice of course type because students choosing telecourses differ from those choosing traditional courses. If choice is not permitted, attrition will most likely be nonrandom, rendering the results uninterpretable.
2. The design should yield an estimate of student gain from the telecourse. The magnitude of this gain should not be indexed to the magnitude of gain in traditional courses because, given population differences, the growth curves may not be the same.
3. The design should provide information about the *extent* of students' knowledge, skills, and attitudes at the end of the telecourse, and how these outcomes compare with those of students in traditional courses. Put another way, the design should be *comparative*; the purpose of the traditional course is to provide a standard against which to evaluate the telecourse outcomes.

EXCHANGEABILITY: A UNIFYING CONCEPT FOR EVALUATING STUDENT OUTCOMES

The central question that an evaluation of student telecourse outcomes must address is: "Are the knowledge, skills, and attitudes acquired by students in a telecourse *exchangeable* for the knowledge, skills, and attitudes acquired by students in an equivalent traditional course?" Exchangeability refers to the extent to which the knowledge, skills, and attitudes acquired by students from a telecourse are interchangeable with the knowledge, skills, and attitudes that are: (a) valued by faculty and administrators, and (b) acquired by students enrolled in the same course offered as part of the traditional curriculum.

The concept of exchangeability has three noteworthy attributes. First, it reflects questions raised by policymakers: "Do students learn from the telecourse what students in a traditional course learn?"

Second, exchangeability introduces a new perspective on the old axiom derived from past media comparison research: "media don't make a difference" in achievement.

[O]verall, there is no basis in the research for saying that students learn more or less from television than from classroom teaching. This does not mean that under some conditions of teaching some students do not learn more of certain subject matter or skills from one medium or channel of teaching than from the other. But the results of the broad comparisons say that there is, in general, *no significant difference*. (Schramm, 1977, p. 28)

In general, media comparisons and studies of the instructional impact of a new medium have yielded little that warrants optimism. Even in the few cases where dramatic changes in achievement or ability were found to result from the introduction of a medium such as television, as was the case in El Salvador (Schramm, 1977), it was not the medium per se which caused the change but rather the curricular reform which its introduction enabled. (Clark and Salomon, in press, p. 10)

If media "don't make a difference," there is good reason to believe that the answer to policymakers' exchangeability question is, "yes," alternative courses are exchangeable. But this past research has only addressed the question of exchangeability for students drawn from the *same* population (either by random assignment or (inadequate) statistical adjustment). For populations nested in tele- and traditional courses, the answer might differ.

Third, exchangeability translates directly into a set of questions that guide evaluation: (a) Does the telecourse content meet academic standards and, if so, do the students acquire the knowledge and skills imparted? (b) Do students in a telecourse learn as much as students in traditional courses that cover the same objectives, assignments, and materials? and (c) Do students in a telecourse learn as much as students in another course that fulfills the same curricular goals but does not necessarily use the same assignments and materials?

Several corollaries follow from the exchangeability notion. The first is that "treatment" and "control" groups in an evaluation—the telecourse and traditional course—are defined as the *curriculum* and its *local implementation*. By local implementation we mean instructor quality, course requirements and credit, and institutional facilities and support. Unlike a laboratory experiment, no attempt is made to disentangle the complex that comprises the "course." For this reason, almost as much interest attaches to variation in implementation and effects *within* telecourses and traditional courses as between them.

The second corollary is that student characteristics and self-selection into tele- and traditional courses cannot be disentangled; they are part of local implementation. Consequently, an evaluation should not attempt to "equate" students in telecourses and traditional courses, either by design (e.g., random assignment, matched assignment) or by

statistical methods. Exchangeability focuses attention on the level of knowledge and skills attained by the end of the course, not the magnitude or "growth" of knowledge and skills in the two courses from beginning to end. That is, students in one kind of course might have gained more knowledge and skills than students in another course, but the level of knowledge and skills attained by students in the second course might be greater than students in the first course. Exchangeability looks at the latter outcome, although it treats the former as a necessary, important condition.

The third corollary of exchangeability is that an evaluation should use designs that balance internal validity (correct attribution of causes for observed tele-/traditional course differences) and external validity (inference to the larger set of telecourses and traditional courses not observed in the study). For example, a randomized experiment might be preferred on internal validity grounds, but not on external validity grounds. Students are seldom, if at all, randomly assigned to a telecourse or traditional course; they typically self-select into one or the other.

IMPLEMENTATION FEASIBILITY

The choice of a design to evaluate student telecourse outcomes also depends on the feasibility of implementing alternatives. A number of institutional factors come into play. Nowhere is there more academic freedom in what and how to teach than in institutions of higher education. For this reason, the greater the control over the educational setting required by a design, the more difficult it may be to implement. Adherence to the planned treatment conditions will be difficult to maintain, especially if problems arise with the instructional approach. And instrumentation will be difficult to protect, particularly if the evaluation team, and not the instructor, prepares the examinations. These problems are exacerbated by the necessity of most telecourses to support themselves, and by the fact that, often, telecourse instructors are not regular faculty members and their employment may depend on course enrollments.

There are two implications for the choice of evaluation design. On feasibility and student-choice grounds, random assignment of students and/or instructors to tele- or traditional course will probably (a) change the nature of the treatments, (b) change the nature of the student population, or (c) both. The one exception might be the very rare institution that cannot offer enough traditional sections of a course to satisfy demand and, consequently, provides a telecourse to the

overflow. Second, the design must include an implementation study, one in which the case-study evaluator is actively involved in the evaluation on-site, attempting to control as well as research the implementation of the evaluation design.

DESIGN RECOMMENDATIONS

Once randomized designs have been ruled out, three designs can be identified as relevant to evaluating student outcomes from telecourses.

Uncontrolled Assignment to Form Non-Equivalent Groups

In Design 1, *Uncontrolled Assignment to Form Non-Equivalent Groups*, students self-select into tele- or traditional courses. Before and after treatments (local implementation of tele- and traditional courses), knowledge, skills, and attitudes are measured. This design compares intact classes,¹ and recognizes that student populations cannot be separated from course implementation. For example, the reasons students choose a telecourse (convenience and accessibility) are different from the reasons students choose a traditional course, thus leading to differences in the nature of the course and the students served. A disadvantage of the design is selection bias—the observed differences arose because of prior existing differences between the populations of students taking the telecourse and the traditional course. This is problematic if we are concerned with (a) the performance of students from the *same* population in the two courses, or (b) the relative *gain* in knowledge and skills between the two groups from the beginning to the end of their respective courses. The very motivation for telecourses (i.e., to serve new student populations) suggests to us that the question of the performance of the *same* population is of little interest in most instances. Gain in knowledge and skills is important as a prerequisite to interpreting the more important *policy* question—to what extent are the *outcomes* of tele- and traditional courses exchangeable?

The Patched Up Design

Design 2, the *Patched-Up Design*, is appropriate when institutions regularly cycle students through the same course. Students from one cycle serve as a control group for students in another cycle. This

¹The use of intact classes raises issues about the appropriate unit of analysis (see Burstein, 1980).

design requires that the institution offer the same telecourse in consecutive (fall and spring) semesters of the same year, and that pretest achievement (and other) data obtained in one semester overlap significantly with the data collected in the second semester (e.g., the spring pretest should contain a substantial number of items from the fall posttest, and the fall posttest should be given as the spring posttest as well). Comparisons of fall posttest data (treatment group) with spring pretest data (control group) would help answer the question of whether students learn more from telecourses than they would have learned without it. Spring pretest-to-posttest achievement comparisons provide an opportunity to replicate the "treatment effect" and increase confidence that some unidentified factor caused the change in performance. The major advantage of this design is that it is less obtrusive than Design 1—existing tests are used in regularly scheduled courses.

Case Study Methods

Finally, *Case Study Methods*—narrative accounts of an object of social inquiry, such as a classroom, a school system, or other bounded system—focus on the cultural context and are usually more descriptive than conclusive. This method obviates some of the problems with experimental designs, such as ignoring important variables (e.g., quality of program presentation, organization and context) or not accounting for others (e.g., differences between individuals or the quality of treatment and control conditions). Case study methods are important for understanding tele- and regular course outcomes. However, these methods alone are unlikely to resolve questions about exchangeability; they should be used in conjunction with other designs. Ultimately, policymakers demand quantitative achievement data to answer potential critics.

INAPPROPRIATE DESIGNS

The following designs are considered inappropriate for evaluating student outcomes from telecourses.

Controlled Assignment to Form Non-Equivalent Groups

Evaluators might use controlled assignment to form non-equivalent groups in two situations. The first is to ensure that the alternative groups represent populations that differ in known ways. For example, the composition of telecourse and traditional course populations may

be known to differ in age, degree objective, and previous education. Rather than leave student enrollment to chance, evaluators might assign students according to those measured characteristics. We believe that this design is *inappropriate* when evaluators' interest is in comparing *equivalent* or *non-equivalent* groups. In the former case, the non-equivalence cannot be analyzed away, even with covariance or related techniques. In the latter case, when evaluators are willing to accept non-equivalent populations, the design has the same problems as controlled assignment (mortality, etc.) and none of the benefits. When evaluators accept the populations as they exist, they assume that the students who enroll in the telecourse and traditional course represent their respective populations.

A special case is the regression discontinuity design (Campbell and Stanley, 1963; Cook and Campbell, 1979). Students are assigned to educational treatments on the basis of a cutoff score on some student characteristic (or a combination of student characteristics). Students above and below the cutoff score are assigned to different treatments. This design is used when allocating a scarce good to the most needy or the most qualified. In compensatory education, for example, students scoring below a certain point on an achievement test might be assigned to remedial education, whereas students scoring above that point might be assigned to regular education (a control group). When regression lines of posttest on pretest are computed separately for each group, a discontinuity appearing at the cutoff point indicates a difference between treatment and comparison conditions. Because a telecourse is not a scarce good, it makes no sense to restrict enrollment to a prespecified population.

Multivariate Nonrandomized Matching Design

Sherwood, Morris, and Sherwood (1975) have described an alternative to randomized experiments when the goal is to compare equivalent groups: a multivariate nonrandomized matching design. A matching design recognizes that treatment and control populations may differ and consequently analyzes only data from those subjects that can be matched on prespecified characteristics. These "matched" groups are then considered equivalent. Sherwood et al.'s variation on the matching design measures a large number of variables prior to the study and uses multivariate techniques to match subjects in the treatment and comparison groups.

Aside from the problems inherent in matching of any kind (which Sherwood et al. discuss), this design is inappropriate for telecourse evaluations for two reasons. First, students in the two populations

differ on so many variables that true matching is unlikely. Second, the overlap between groups may be too small to produce sufficiently large samples for analysis. For example, if students enrolling in a traditional course tend to be full-time undergraduates intending to complete a B.A. degree, whereas students in a telecourse tend to be older, part-time students not seeking a degree, the overlap between populations may be too small to produce matched samples of reasonable size.

III. A TALE OF TWO TELECOURSES

We next describe the two telecourses that provided the testbed for our feasibility study: *The Constitution: That Delicate Balance* and *The New Literacy* (an introduction to computers). We describe the courses both as they were designed and as they were implemented at specific sites, characterize the students who took the courses at each site and their course participation, and identify characteristics of the course implementation and the sample that created difficulties for the evaluation process. The variability in the implementation of both courses demonstrates the necessity of embedding analysis of student outcomes in richly textured descriptions of implementation processes.

Both courses are part of a series of telecourses recently created by the Annenberg/Corporation for Public Broadcasting (A/CPB) Project. Established in 1981 by a grant of \$150 million from the Annenberg School, the A/CPB Project's mission is to demonstrate how telecommunication systems can reach new audiences in higher education, and to develop innovative collections of high-quality, college-level courses to facilitate "distance teaching." The series covers diverse topics: the brain, Congress, physics, written composition, computers, and the Constitution. The series uses high-quality production in an attempt to match the expectations of today's students, who are sophisticated video audiences. We decided to focus on *The Constitution* and *The New Literacy* because, although they differ as much in production style as they do in content, they appear representative of the highest quality telecourses available today. Both use the visual medium to extend the instructional repertoire available to the traditional classroom teacher.

DATA GATHERING PROCEDURES AND INSTRUMENTS

We gathered data on the two courses at each site in several ways, including administering both pre- and posttests and a postcourse questionnaire.

The pretest for *The Constitution* sites consisted of a background questionnaire, a verbal reasoning test, and an 18-item achievement test. The background questionnaire provided information on students' demographic and academic backgrounds, television-viewing habits, study habits, and attitudes toward telecourses. The verbal reasoning

test measured general intellectual ability.¹ The achievement pretest used 18 of the 100 multiple-choice items that the research staff developed for the midterm and final.

The pretest for *The New Literacy* contained the same background questionnaire and verbal reasoning test, and an achievement test. For reasons explained below, we administered two forms of the achievement test at the sites offering this telecourse.

The posttest for each course constituted the final exam. At the same time, a postcourse questionnaire was administered that provided information on the students' course participation ("studenting"), perceptions, and attitudes.

To gain information on the implementation of the course at each site, we interviewed instructors and administrators before, during, and after the course. Members of the research team attended and monitored pretests, midterm examinations, and posttests at most sites.

We designed our evaluation procedures to be as unintrusive as possible; however, some of the procedures, such as our input into testing, clearly changed the implementation of the telecourse at the sites (except when an instructor chose not to cooperate).

OVERVIEW OF IMPLEMENTATION

Implementation for both telecourses was highly variable and demonstrates the necessity of embedding analysis of student outcomes in richly textured descriptions of implementation processes. Variability characterizes the way the courses were implemented, the characteristics of students served and their course participation, and the problems and difficulties encountered by the evaluators. These results support the notion that each telecourse "treatment" is different and is defined by the telecourse and its local implementation. Below, we discuss each course in detail.

THE CONSTITUTION

As designed and distributed, *The Constitution: That Delicate Balance* is a one-semester introduction to constitutional law and public policy. Course materials include both video and traditional media: 13 one-hour videolessons that are broadcast locally, a study guide, and a

¹We used the Extended Range Vocabulary Test from the Educational Testing Service's *Kit of Reference Tests for Cognitive Factors* (French, Ekstrom, and Price, 1963). The test has 24 items, takes about six minutes to complete, and is appropriate for grades 7 through 16. The test measures knowledge of word meanings.

textbook.² The textbook for *The Constitution* is written in a chatty, popular style and overlaps only intermittently with the videolessons. The study guide is larger, more comprehensive, more scholarly, and more insistently instructive.

We examined the implementation of the *The Constitution* at three campuses, which we will refer to as Sites A, B, and C. The implementation of the telecourse varied substantially across these sites, although not so greatly as across *The New Literacy* sites we report on below. Some of these differences in course implementation created difficulties for our evaluation.

Variability in Course Implementation

Table 1 summarizes the characteristics of the implementation of *The Constitution* at the three sites we studied. The implementation varied widely on many dimensions.

Sites A and C are universities; B is a community college. At each site the telecourse was offered for comparable credits, although at Site A the course cost students approximately ten times as much as at other sites, possibly providing extra incentive for these students to finish the course.

Course format and requirements varied among the three sites. All required an orientation session and met between five and ten times during the term for discussions and examinations. All required viewing the videolessons and reading of the McKenna text. However, the instructor at Site B administered four rather than two examinations and did not require reading of the Friendly and Elliot text. The instructor at Site A supplemented the midterm and final examinations with other evaluative instruments: class participation and a paper. Instructors' grading practices were also variable; Site C based grades on examinations only, whereas Site A placed more weight on a written paper and class participation.

The instructors for the three courses all rated the materials for the telecourse highly, but varied in their experience in teaching telecourses and in their understanding of their role in the telecourse. The instructor at Site A was a part-timer with much prior experience teaching both regular courses and telecourses (although he much preferred telecourses), and a strong reputation as a superior teacher. He viewed his role as keeping student debate on track, as well as updating students on judicial decisions and legislative policies that related to telecourse materials.

²The study guide is G. McKenna, *A Guide to the Constitution: That Delicate Balance*, Random House, New York, 1984. The textbook is F. W. Friendly and M. J. Elliot, *The Constitution: That Delicate Balance*, Random House, New York, 1984.

Table 1
TELECOURSE IMPLEMENTATION AT
THE CONSTITUTION SITES

Characteristic	Site		
	A	B	C
College type ^a	U	CC	U
No. of semester units	3	4 ^b	3
Course work:			
View all videolessons	x	x	x
Read McKenna text	x	x	x
Read Friendly text	x		x
Number of examinations	2 ^c	4 ^d	2 ^e
Graded class participation	x ^f		
Graded term paper	x ^f		
Class meeting schedule:			
No. of on-campus meetings	6	9	8
Orientation	x	x	x
No. of meetings for discussion only	3	4	5
No. of meetings for discussion and examinations	2	0	0
No. of meetings for examinations only	0	4	2

^aU = University; CC = Community College.

^bQuarter units.

^cEach examination had 50 objective items and one essay question and accounted for 20 percent of the final grade.

^dEach examination had 25 objective items and one essay question developed by the study team.

^eStudents answered six of 18 essay questions for each examination.

^fAccounted for 20 percent of the final grade.

^gAccounted for 40 percent of the final grade.

At Site B the instructor was a full-time faculty member and department chair who was teaching his first telecourse. Although he considered the telecourse materials first-rate, this instructor was displeased with the role that the telecourse format imposed on him. Interchange among students and between teachers and students, he argued, was the essence of teaching; unfortunately, it was the very nature of telecourses to view the instructor's role in promoting interchange as superfluous. Thus, he enjoyed teaching the telecourse far less than his traditional courses; he was not there to teach but to "manage," and this he considered a misuse of his energy and ability.

The instructor at Site C was an able, experienced full-time faculty member who was teaching his first telecourse. He was highly

enthusiastic about *The Constitution*, which he conceived as a course for American citizens on "public policy and public ethics" rather than on constitutional law.

Evaluation Implementation at *The Constitution* Sites

Because our intent was to study the feasibility of evaluating telecourses, the problems we encountered in implementing our research are of special interest. They are summarized in Table 2. We encountered some problems at all sites, although only those at Site C, where the instructor did not properly administer our instruments, actually undermined the evaluation effort. Two important problems were common to all three sites: (1) small sample size and (2) the intrusive nature of evaluation instruments with objective items in a course where instructors ordinarily prefer short answer and essay exams.

Site A. At Site A we believe that strong administrative support, an unusually able and experienced instructor, and an on-site, academic representative of our project to mediate potential sources of strain all contributed to efficient data collection. The instructor allowed us to alter his customary testing procedures in several ways. First, he allowed us to substitute our pretest for his own. More significantly, although he preferred essay and short answer tests to objective tests, he cooperated fully in giving and grading students on our objective tests (supplemented by an essay question).

Implementation problems with our tests were few. Some students did not have enough time to complete the pretest and so took it home and mailed it to our on-site monitor. Apparently the students were not happy with all of our measures, but were "good sports" and completed them anyway.

The instructor objected to the length of our 55-question midterm, so we reduced the final examination to 40 questions. Finally, although students did complete our posttest questionnaire in class, they found it burdensome to do so because, in addition to the final examination, they had to present abbreviated oral versions of their term papers and to fill out an in-house course evaluation form. Thus, there were some sources of aggravation in testing, but the instructor remained highly cooperative and was particularly happy to have an on-site monitor to mediate whatever problems did arise.

Site B. At Site B, implementation of the research design was free of conflict but there were a few procedural hitches. Administration of the pretest and posttest proved to be somewhat disruptive. Students complained about doing them for an anonymous research group (particularly as they had not been previously informed about this added

Table 2
EVALUATION IMPLEMENTATION PROBLEMS AT
THE CONSTITUTION SITES

Characteristic	Site		
	A	B	C
Sample size:			
Low enrollments forced two course cancellations	x		
Low enrollment almost caused course cancellation		x	
Site was added mid-semester to increase sample size			x
Pretesting:			
Students finish the class, but do not complete pretest ^a	x	x	x
Attrition:			
Students take the pretest but do not finish the class ^b	x	x	x
Instrumentation:			
Objective items were very different from essay/short answer tests customarily given	x	x	x
Data were returned very late after spring semester			x
Instructor:			
Critical of telecourse			x ^c
Critical of instrumentation	x ^d		
Administered the Rand items after a three-hour essay examination			x
Asked students to complete questionnaires at home and mail them			x

^aSite A did not pretest two students; Site B did not pretest four.

^bSite A lost two students; Site B lost four.

^cInstructor thought his role more "managing" than teaching.

^dInstructor wanted to review examinations before review session and requested that the test be shortened.

course "requirement"). The instructor found long-distance communication with the researchers alien to his routines; he was far more comfortable when the staff met with him in person. It was difficult to contact him by phone to review test questions and to get updates on the progress of the course. Moreover, the inconvenience of photocopying and mailing exam copies caused the instructor to delay turning in test data from three of his tests until we picked them up at semester's end. In brief, there was considerable inefficiency in gathering data.

Site C. At Site C, the evaluation was complicated because the telecourse at this site began unusually early in the year, before we were able to prepare a pretest. We decided to include Site C in the study only later, with the sole purpose of obtaining additional item data once it became clear that final enrollments at the other two study sites would be low. By then, Site C students had already taken a midterm

essay examination. We arranged with both the telecourse administrator and the instructor to give our posttest questionnaire and a 50-item final examination, not to the exclusion of but to substitute for several of the essay questions.

To the surprise of both the telecourse administrator and the researchers, the instructor gave our final examination *after* his own three-hour, six-essay final examination. We suspect that student exhaustion and the fact that the test was irrelevant to their grades compromised the item data. In addition, posttest questionnaire returns were disappointing because the instructor asked students to fill the questionnaire out at home and return it by mail.

Student Characteristics

Data on student characteristics at Sites A and B indicate that telecourse populations are not homogeneous—multiple populations are served depending on the site. (Because Site C was recruited mid-semester, the pretest was not administered there and information on student characteristics is not available.) Although the sample sizes were small ($N = 8$ and $N = 5$ at Sites A and B, respectively), the results indicate that students are quite different between sites. Students at Site B tended to be younger, work less, and have higher incomes and educational attainment than students at Site A. At Site A, students were more often minorities and fewer had full-time jobs. Further, they tended to have lower verbal aptitudes and attitudes toward telecourses, and less telecourse experience. A substantially smaller percentage was enrolled in a degree program. Half at each site had taken one or more government courses. However, a greater percentage of students at Site B: (a) were not enrolled in a degree program, (b) would be willing to take the same course as a traditional course on campus, and (c) had not taken a telecourse before. Interest in the course was equally high at both sites.

These between-site differences raise a question: Should the courses be offered differently to accommodate the needs of different types of students? Future research might examine the relationship between who is served and the curriculum presented.

The dropout rates at Sites A and B were 20 and 44 percent, respectively. Although we could not determine *why* students drop out, our data indicate that compared with students who attended orientation and finished the telecourse, dropouts who attended orientation tended to be younger, less committed (as evidenced by lower enrollment in a degree program or in a credit-gaining course), and lower in verbal abil-

ity and prior knowledge about the Constitution. Future research might examine the dropout issue more closely.

Course Participation

At the conclusion of the telecourse, students at all three sites answered questions regarding their participation or "studenting" in the course, in addition to taking an achievement test. Their responses are summarized in Table 3.

Several findings highlight, once again, differences between sites. First, students at Site B reported, on average, attending fewer meetings, getting less out of those meetings, and studying less than did students at Site A or Site C. Second, at Site B the video may have played a considerably smaller role in the course than at Sites A and C. None of the students at Site B reported taking notes while viewing the telecourse, whereas roughly three-fourths or more of the students did at the other two sites. The explanation for this finding is that

Table 3
COURSE PARTICIPATION AT
THE CONSTITUTION SITES

Characteristic	Site		
	A	B	C
	Sample Size		
	10	7	14
Mean:			
No. of on-campus meetings attended	4.7	2.7	4.1
Study hours per week	3.7	2.3	3.1
Rating: ^a			
Helpfulness of meetings in learning	4.1	2.3	3.9 ^b
Instructor preparation	4.7	3.9	4.0 ^b
Percent:			
Reading >65% of McKenna textbook	90	86	100
Discussing TV program with others at end	80	57	79
Learning			
More from video	0	0	0
More from text	50	86	40
Equally	50	14	60
Taking notes while reading	70	86	86
Not contacting instructor	60	29	50

^aScale of 1 (low) to 5 (high).

^bSite A significantly greater than Site B.

students at Site B were told at a general orientation *not* to take notes on the video, whereas students at the other two sites were either told to take notes or told they would be held responsible for the content conveyed in the video portion of the course. In addition, fewer students at Site B discussed the program following its viewing, and a smaller percentage considered the video and text equally beneficial in learning the material. In contrast, substantial percentages of students at all sites reported use of text material. Nevertheless, telecourse and textbooks were both rated uniformly high by students at all three sites. Third, although not statistically significant, the instructor at Site A was rated, on average, higher than the instructors at the other two sites. This finding is consistent with implementation study findings. And fourth, a substantially larger percentage of students at Site B contacted their instructor for assistance.

In sum, "studenting" varied across sites in terms of course requirements, time students spent studying, what students studied, and so on.

THE NEW LITERACY

The New Literacy is a one-semester introduction to computers. Twenty-six locally broadcast, half-hour programs were created out of the textbook and a study guide was developed to integrate the videolessons with the text.³ In four of the five telecourse sites for *The New Literacy*, the instructors did little more than advise students at the beginning of the semester to watch the videolessons, and did not refer to them again. The textbook drove the course, supplemented to a far lesser extent by the telecourse study guide. Only the telecourse instructor at Site H, who met his class weekly, referred at all regularly to the videolessons. And he did so far less after the first few weeks when, in his judgment (and, apparently, that of his students), the videolessons became repetitive.

Variability in Course Implementation

We examined implementation procedures for *The New Literacy* at five sites. Table 4 summarizes the telecourse implementation characteristics. Of special interest is Site D because it offered in addition to *The New Literacy* telecourse (D1), a traditional introductory computer course (D2) equivalent in objectives, credits, and required texts.

³The textbook was M. Bohl, *Information Processing*, 4th ed., Science Research Associates, Chicago, 1984; the study guide, M. Bohl, *Telecourse Guide for the New Literacy*, Science Research Associates, Chicago, 1984.

Table 4
TELECOURSE IMPLEMENTATION AT
THE NEW LITERACY SITES

Characteristic	Site					
	D1	D2	E	F	G	H
	Sample Size (Pretest)					
No. of semester units	47	18	44	67	33	27
Course work:	3	3	3 ^a	2	2	3
View all videolessons	x		x	x	x	x
Read entire text	x	x	x	x	x	x
Read entire guide	x		x	x	x	x
No. of examinations	2	2	2	2	2	2
No. of quizzes	6 ^b	0	0	0	0	2
Extra-credit questions and projects	x					
Computer programs				x		x
Final paper		x				
Class meeting schedule:						
No. of on-campus meetings	5	18	5	4	5	18
Orientation	x		x	x	x	
Study planning	x					
Midterm review	x		x		x	
Midterm examination ^c	x		x	x	x	
Two field trips						x
Final review	x		x	x	x	
Final examination ^c	x	x	x	x	x	x

^aQuarter units.

^bTake home.

^c100-item, multiple-choice examinations were developed for Sites D1 and D2. A subset of these items (15 core items) was added to Sites E through H's examinations.

However, the two courses had different instructors, and administrative authority over the two courses was entirely separate. *The New Literacy* was run by a unit that administered only telecourse and independent study programs, whereas the traditional course was under the authority of a regular college department. The four sites in addition to D1 and D2 are E, F, G, and H. All are community colleges.

Course format and requirements for *The New Literacy* telecourse varied widely at the five sites. All required an orientation and met about five times, except for Site H, which met as frequently as the traditional course at Site D2 (18 meetings). All telecourse sites

required viewing the videolessons, reading the textbooks, and taking midterm and final examinations. Site D1's instructor also required students to complete six quizzes and return the answers on quiz cards by dates specified throughout the semester. Within a week, students received prescriptive feedback indicating which questions were incorrectly answered and where in the text the correct answers could be found. In addition, students could obtain extra credit by completing two two-page miniprojects from the telecourse study guide or attending a special workshop. At Site H, students took six in-class quizzes on text assignments and field trips. The instructor at Site D2 substituted a final project for a final examination in determining students' grades. Not surprisingly, this variety in course requirements resulted in a variety of grading practices.

The instructors in *The New Literacy* courses and the administrative climate in which they taught were all very different. At Site D, telecourses have been under attack for some time, from both local faculty and administrators and faculty of the state university system who regulate transfer of community college course credits into four-year institutions. The former fear loss of already shrinking community college enrollments to telecourses and part-time instructors, whereas the latter contend that telecourses generally do not conform to university "standards." In response, Site D has upgraded telecourse requirements by incorporating a required written composition component in addition to regularly scheduled examinations. These changes have not yet improved the status of telecourses in the eyes of critics, even though telecourse requirements are often more stringent than in traditional courses. In addition, *The New Literacy* was run by a unit that administered only telecourse and independent study programs, whereas the traditional course was under the authority of a regular college department. The telecourse instructor was a part-timer who had taught the telecourse before and had participated in its development. The traditional course instructor was new to community college teaching and was teaching the course for the first time. Both instructors had full-time jobs in computer-related fields.

At Site F, high-level administrative support for our study was lukewarm, as the college has traditionally been guarded in allowing outside evaluators on campus. Moreover, the telecourse administrator felt that *The New Literacy* was not a particularly innovative use of television as an instructional medium. He also felt that the pretest asked far too much of students without paying them to complete it. The instructor at Site F was a part-timer who had declared before the course began that he was philosophically opposed to telecourses, had

not viewed all *The New Literacy* videolessons, and thought the telecourse should receive only one college credit.

At Site G, *The New Literacy* was conducted professionally as standard operating procedure. The telecourse was offered as a regular department course and was taught by a full-time faculty member. Telecourse delivery at Site G was facilitated by the highly enthusiastic support of an experienced, high-ranking administrator who, while deferential to departmental prerogatives in deciding which telecourses to offer and how to staff them, exerted considerable power on his own. Moreover, he enjoyed active support for telecourses from the campus president and vice-president.

Administrative practices at Site H mandated that all full-credit courses, including telecourses, hold weekly meetings. The only structural difference was that a traditional course met for three hours, whereas a telecourse met for one and one-half hours. Thus, Site H's telecourse, with its weekly meeting structure, was in stark contrast to other sites. The instructor was a part-timer whose primary employment required considerable travel, and whose principal contact was his department supervisor. The lack of communication between the instructor and telecourse administrator caused data collection problems for the evaluation.

Finally, Site E's instructor was a part-time last-minute replacement who was teaching his first telecourse (although he had taught many traditional computer courses). Again, miscommunications between instructor and administrator caused confusion—the instructor did not know that we would supply the examinations when he agreed to teach the course. Further constraints posed by the administrator concerning course orientation and pretesting procedures resulted in a greatly reduced pretest sample at this site.

Evaluation Implementation at *The New Literacy* Sites

Problems encountered in our evaluation of *The New Literacy* are summarized in Table 5. Only Site G was without major problems. We describe samples of the problems encountered at the other sites below.

Site D. The chief difficulty at Site D was the examinations, which were developed by the Rand research team. Many of the items were drawn from an existing pool of objective test items developed specifically for this telecourse and used by nearly every *New Literacy* research site.⁴ Although Site D instructors agreed that we could provide the test, had reviewed both examinations, and had input into their revision,

⁴Items were created by the Southern California Consortium for Community Colleges with assistance of Interwest Applied Research.

Table 5
EVALUATION IMPLEMENTATION PROBLEMS AT
THE NEW LITERACY SITES

Characteristic	Site					
	D1	D2	E	F	G	H
Pretesting:						
Students finish the class, but do not complete pretest ^a	x	x	x	x	x	x
Attrition:						
Students take pretest, but do not finish the class ^b	x	x	x	x	x	x
Data:						
Unavailable for fall semester						x
Returned very late after spring semester						x
Instructor:						
Critical of telecourse	x ^c				x ^d	
Critical of instrumentation	x ^e	x ^f	x ^e	x ^e		
Told students that the final examination did not affect their grade		x				
Gave students a comprehensive outline of final examination items	x					
Gave students precise answers to midterm and final exam items				x		

^aSites D1, D2, E, F, G, and H did not pretest 29, 1, 45, 11, 6, and 6 students, respectively.

^bSites D1, D2, E, F, G, and H lost 18, 15, 15, 35, 17, and 11 students, respectively.

^cOverload of information and overspecificity.

^dPhilosophically opposed to telecourses.

^eToo lengthy and difficult.

^fThe instructor felt the evaluator had undermined his prerogative and antagonized students against the course.

^gDisagreed with item answers.

both expressed dissatisfaction with them in the end. Student complaints following the midterm examination prompted the instructor at Site D1 to disavow any responsibility for the exam and to consider grading the midterm on a curve despite college policies against such grading practices. For the final examination review, the instructor prepared a comprehensive outline of each chapter which posed key questions and identified relevant page numbers for answering them.

Implementation of research procedures in the traditional course, Site D2 ("control group"), turned out to be far more difficult than anticipated and fundamentally subverted the potential value of Site D as an evaluation site. Starting with an enrollment of 18, the course ended with only three students, far too few to produce conclusions about

exchangeability of student outcomes. Moreover, the instructor peremptorily altered the course format—against our urgent request that he not do so—by substituting an individual student project for our final examination as a course requirement. The three students who completed the course did, however, take our final examination and apparently did so seriously, even though it did not figure in their grades.

In retrospect, these substantial changes appear to have derived, in part, from the examinations and research study procedures. The instructor at Site D2 felt that the testing procedures had undermined his relations with his students by making them "fearful." Although the instructor was given advance opportunity to challenge midterm and final examination questions, he later asserted that the specificity of the tests controverted the course's intent, that, indeed, the tests contradicted an agreement he had made with students at the first class meeting that he would focus on basic computer concepts rather than details. He also felt that many test items were poorly written, ambiguous, and open to challenge. These problems suggest a warning that future evaluations should heed: an outside evaluator's involvement in the actual conduct of college courses is intrusive and anomalous and may be perceived as undermining instructor autonomy and control over students.

Site E. Although overt problems were less evident at Site E, the study design was subverted there more thoroughly than at any other site. Problems first arose at the pretest administration. In spite of prior arrangements, the pretest was administered after the mandatory orientation session, in a separate location, and without prior knowledge on the part of students. As a result, the pretest sample was much smaller than desired due largely to the administrator's decision not to incorporate the pretest into the orientation session itself.

Although this did not unduly harm the research design, the instructor totally undermined the research effort during the midterm and final review by reading the examination questions in order and providing the right answers to each one. Not surprisingly, students in *The New Literacy* at Site E did fantastically well on their examinations compared with students taking the course the previous semester (fall 1984).

Site F. At the final examination, without warning, Site F's instructor chose not to use our test. After altering the position and adding alternative responses to several of the common test items, the instructor decided at the last minute to restore the common items to their original form. These last-minute changes caused the instructor to run out of time to get the examination itself ready for distribution. In the end, a researcher was obliged to staple the tests together while the instructor conducted an impromptu review.

Site H. A key implementation problem from the start (although we did not learn about it until much later), was the minimal contact between the administrator, nominally responsible for telecourses, and the course instructor. Site H was included in the study only after the administrator had assured us that the instructor (who had also taught *The New Literacy* in fall 1984), had retained midterm and final examination data essential to the patched-up design. This turned out (we learned halfway through the semester) not to be the case. In the end, the only item data usable from Site H were the comparison questions common to the pretest and final examinations: the 15 common items administered at pretest, midterm, and final examination sessions, and the pretest and final student questionnaires—far less data than from any other site.

Student Characteristics

Student demographic characteristics are provided in Table 6. (Although we table data from Site D2, the small sample size precludes our interpreting them.) The data in this table indicate that the student populations served by *The New Literacy* systematically differ in age, percentage of women in the course, and in percentage of students holding bachelor's degrees. Sites H and D1 represent the boundaries on age and gender. Site E contains proportionately more and Site F proportionately fewer students with bachelor's degrees.

The sites also systematically vary in average ability but not in attitudes. Sites H and D1 provide the bounds on mean verbal ability and prior computer knowledge. The other three sites are roughly similar, with Site E higher in verbal ability and Site G higher in prior knowledge.

In contrast, students at all five sites are similar in reported grades, enrollment in degree programs, enrollment in the course for credit, computer-course background, and interest in the course. Proportionately more students at Site H than at other sites say they would enroll in a traditional course if the telecourse were not offered, and Sites E and D1 represent the boundaries of previous telecourse experience.

Although finishers tend to look different from dropouts, only the differences for aptitude and achievement are consistently statistically significant. Finishers tend to be, on average, higher in verbal ability (except at Site F) and in prior computer knowledge. In addition, at Sites H and D1, there were more minority student dropouts than finishers. Finishers and dropouts did not differ significantly on any other demographic characteristic or on educational background. Again, our data cannot determine *why* students drop out.

Table 6
STUDENT CHARACTERISTICS AT THE NEW LITERACY SITES

Characteristic	Site						p < .05 Between Sites
	D1	D2	E	F	G	H	
	Sample Size ^a						
	29	2	29	32	16	15	
Mean:							
Age	40.5	37.0	40.1	32.3	36.1	30.2	x
Verbal ability ^b	15.8	17.0	16.0	13.5	13.3	13.1	
Computer knowledge pretest ^c	7.0	5.0	6.0	5.3	6.6	5.1	
Computer knowledge posttest ^c	10.9	12.0	14.6	8.0 (34)	8.8	7.3 (20)	
Percent:							
Female	69	50	61 (28)	59	31	33	x
White	93	100	75 (28)	63	75	93 (14)	x
Household income > \$30K	57.(28)	50	69 (26)	57 (30)	56	53	
Working full time	66	100	72	84	75	80	
Enrolled for a degree	31	0	27 (26)	53	44	43 (14)	
Have B.A. or higher	17	50	43 (28)	13	25	7	x
B or better GPA	76	50	62	71 (31)	56	60	
Not taken telecourse	52	100	96 (28)	63	81	87	x
Not taken computer course	90	100	69	66	60.(15)	73	

^aNumber of students who completed the course and took the pretest and posttest, except as noted in parentheses.

^b24 items—the Extended Range Vocabulary Test from ETS's *Kit of Reference Tests for Cognitive Factors* (French, Ekstrom, and Price, 1963).

^c15 items—representing knowledge gained from *The New Literacy* course.

Course Participation

Students at all five sites answered questions about their "studenting" in *The New Literacy*. We interpret the data in Table 7 as showing variation in telecourse implementation and important differences in students' study habits. (Again, Site D2's responses are reported for completeness only; they were not included in statistical tests.)

Clearly, *The New Literacy* does not represent a single, homogeneous treatment, especially when implementation considerations are kept in mind. Rather it represents a range of treatments with Site H on one extreme. Students at Site H attended class, on average, significantly more often than did students at any of the other sites, and rated the value of these meetings significantly higher than did students at the other sites. Moreover, students at Site H met outside of class, on average, more often than did students at any of the other sites but reported studying significantly fewer hours!

"Studenting" at the other four sites is quite similar, with unsystematic exceptions. For example, students at D1 tended to spend, on average, considerably more time studying, but a smaller than average percentage discussed the video program after viewing it, and a greater percentage contacted the instructor. Proportionately fewer students at Site F read the text assignments and answered questions in the study guide.

In sum, implementation of the telecourse varied from site to site. Nevertheless, students in most implementations of the telecourse attended class meetings two or three times during the semester, studied about three hours a week, and found the text more helpful than the video in learning the material. The only systematic exception to this pattern was Site H. At all sites, the overall preparation of the instructor and the worth of the telecourse were rated as moderately high, as were the textbook and study guide.

SUMMARY

This overview of the telecourse sites highlights the diversity of telecourse implementation in the study. No two sites were even close to being alike. Course formats and grading requirements varied widely in design and even more in operation; instructors differed dramatically from enthusiastic to disgruntled, from highly professional to egregiously unprofessional. Pat lessons from these myriad implementation experiences are clearly unjustifiable. Indeed, if one grand lesson had to be drawn, it would be to expect extreme variation and unreliability in operationalizing evaluation designs in the American system of higher education.

Table 7
COURSE PARTICIPATION AT THE NEW LITERACY SITES

Characteristic	Site						p < .05
	D1	D2	E	F	G	H	
	Sample Size ^a						
	57	2	61	43	24	16	
Mean:							
No. of on-campus meetings attended	2.0	17.5	2.6	2.5	3.3 (23)	15.5 (15)	x
Study hours per week	4.4	6.5	3.6	3.0 (42)	2.6 (23)	1.9	x
No. of outside meetings with students	0.7 (55)	0.0	1.0 (62)	0.5 (41)	1.0 (23)	2.9	x
Rating: ^b							
Helpfulness of meetings in learning	3.1 (56)	4.0	3.4 (63)	3.3 (11)	3.3	4.3 (15)	x
Percent:							
Reading >65% of textbook assignments	77 (56)	100	75 (64)	37	54	69	x
Answering >65% of questions in guide	49	(c)	55 (64)	35	63	31	
Discussing TV program with others at end	40	(c)	69 (64)	51 (41)	67	56	x
Not contacting instructor	35	(c)	53 (64)	50 (42)	67	63	x

^aStudents completing course; may or may not have taken pretest; exceptions are noted in parentheses.

^bScale from 1 (low) to 5 (high).

(c)Not applicable.

Diversity also characterized the students served and their "student-ing" habits. Taken together, diversity in implementation and the population served supports the view that each telecourse "treatment" is different and must be defined to include both the telecourse *and its local implementation*.

IV. STUDENT OUTCOMES FROM TELECOURSES

In this section, we attempt to answer the question, "How much do students learn from telecourses?" Initially, we set out to implement a number of experimental and quasi-experimental designs for studying the exchangeability of student outcomes in tele- and traditional courses. For each design we considered the evaluation question(s) it might answer and possible threats to internal and external validity (see Sec. II). Although some designs were better than others for determining exchangeability of outcomes, research implementation problems (not methodological limitations) hampered the evaluation effort from the beginning.

For *The Constitution*, lack of sites, insufficient lead time, inability to control assignments of students to tele- and traditional courses, and cancellation of courses due to underenrollments permitted implementation of only a pretest-posttest design at Sites A and B and a posttest-only design at Site C (see Campbell and Stanley, 1963, for details of these designs). Neither of the designs provides direct evidence on the exchangeability of student outcomes from tele- and traditional courses. They did permit us to study the implementation of telecourse evaluation (as described in Sec. III) and the statistical (psychometric) properties of the *Constitution* item pool developed for the evaluation.

For *The New Literacy*, Site D offered equivalent tele- and traditional courses (described in Sec. III). Although we attempted to evaluate exchangeability of student outcomes in Sites D1 and D2, 80 percent of the students attending orientation dropped out of the traditional course. At final count, we were left with a traditional course containing three students, only two of whom had attended orientation. Consequently, the comparison of tele- and traditional courses and the implications for exchangeability are much more problematic and less important than we had hoped.

For these reasons, we limit the following discussion of student outcomes to those evaluated by a "patched-up" design. For *The New Literacy* telecourse we: (a) describe the evaluation design, (b) describe the measures (the tests) and evaluate their psychometric properties, and (c) provide preliminary data on achievement and attitude outcomes. The evidence indicates that *The New Literacy* telecourse, as implemented at four sites, significantly increases students' knowledge about computers, by about the same amount.

PATCHED-UP DESIGN—THE RECURRENT INSTITUTIONAL CYCLE DESIGN

When institutions regularly cycle students through the same course, students from one cycle might be considered as a control group for students from another cycle (Campbell and Stanley, 1963). The "patched-up" design requires that the institution offer the same telecourse in consecutive (fall and spring) semesters of the same year, and that pretest achievement (and other) data obtained in one semester overlap significantly with the data collected in the other semester (e.g., the fall and spring examinations are the same and a spring pretest is given that contains a substantial number of items from the final examination). Comparisons of fall posttest data (treatment group) with spring pretest data (control group) would help answer the question of whether students learn more from a telecourse than they would have learned without it. Spring pretest-to-posttest achievement comparisons provide the opportunity to replicate the "treatment effect" (i.e., the effect of the telecourse) and increase confidence that some unidentified factor did not cause a change in performance. The major advantage of this design is that it is fairly unobtrusive since existing tests are used in regularly scheduled courses.

The results of the patched-up design do not indicate whether what students learned in the telecourse is exchangeable for what is learned in an equivalent traditional course. Nevertheless, the design can be modified to do so by incorporating a traditional course into it.

Instrumentation

We applied the patched-up design at four sites at which *The New Literacy* was offered in both fall semester 1984 and spring semester 1985: Sites E, F, G, and H. Our study began in January 1985, so we compared spring students' pretest scores with the fall semester students' final examination scores—providing a control/treatment comparison. We then compared spring pretest scores with posttest scores to see if the "treatment effect" replicated. Finally, we compared fall posttest scores with spring posttest scores to see if the outcomes were equivalent in both semesters.

At the first three sites, test item data were available on the fall midterm and final examinations. (Midterm and final data are combined in all analyses and scores on these examinations are hereafter referred to as posttest scores or final examination scores.) At Site H, we knew each student's percentage correct on the fall final and the number of items on the examination. Consequently, we could calculate statistics

for the examination but could not estimate reliability without making heroic assumptions that we were unwilling to make. At all four sites, spring pretest data consisted of 50 items selected from the fall examinations, and 15 items developed by the research team (hereafter referred to as Rand common items; see the appendix). Finally, since the same spring data were collected at Site D1, this site was included in all relevant analyses to provide an additional replication of the telecourse effect.

Reliability of the Tests

With the patched-up design, we used the instructors' tests, with the exception of (a) the 15 Rand common items given in the spring at pre- and posttest and (b) the Rand-developed spring final at Site D1. The instructors' examinations are especially important. If they are unreliable, considerable doubt is cast on the utility of the patched-up design. If, however, the tests are sufficiently reliable for making mean comparisons, an important hurdle will be overcome.

Internal consistency reliability for the Rand common pre- and posttest and each site's test are presented in Table 8. Internal consistency coefficients provide a lower bound because they assume homogeneous items and the items on the achievement tests cover heterogeneous material. Moreover, pretest reliabilities on subject matter knowledge

Table 8
**INTERNAL CONSISTENCY OF VOCABULARY
AND ACHIEVEMENT TESTS**

Test	Site				
	E	F	G	H	D1
Rand common					
Pretest ^a	.88	.65	.71	.66	.50
Posttest ^a	.99	.89	.92	.85	.88
Site's pretest ^b	.82	.64	.72	.73	.63
Site's fall final ^c	.72	.73	.25	(d)	(d)
Site's spring final ^c	.96	.87	.77	.91	.93

^a15 items, at all sites.

^b65 items, Sites E, F, G; 63 items, Site H; 26 items, Site D1.

^c100 items.

^dNot available.

^e100 items, Site E; 162 items, Site F; 110 items, Site G; 130 items, Site H; 200 items, Site D1.

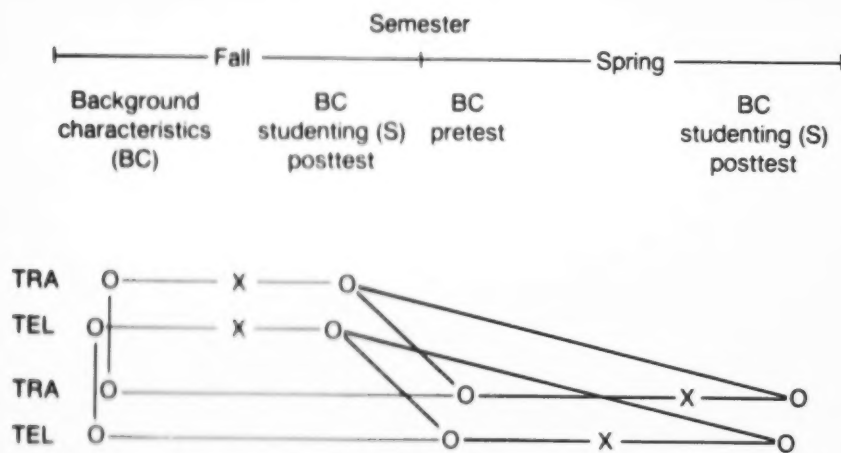
are expected to be lower than other coefficients because students guess on more items.

With this preamble, Table 8 provides evidence that the reliabilities of instructor-made tests are high enough to be used in making mean comparisons. (The one exception is Site G's fall final, which we cannot explain.) Moreover, an inspection of the items on each of the tests led us to conclude that the tests contained representative samples of the content covered in the text. Since the instructors emphasized the text, few items tested information presented in the videolessons only. The patched-up design, then, probably meets minimal psychometric requirements for exchangeability evaluations.

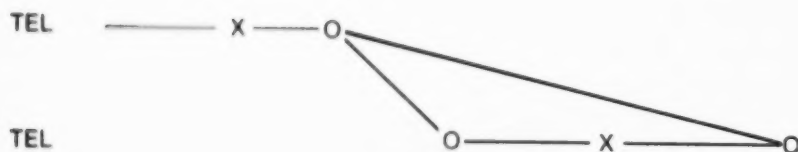
STUDENT OUTCOMES

A schematic of the patched-up design is presented in Fig. 1. A "preferable" version of the design compares a tele- and a traditional course, and provides student background, course participation ("studying"), and test information for both the fall and spring semesters (Fig. 1(a)). We implemented only part of the "preferred" design, as shown in Fig. 1(b). Comparison of the fall posttest measure with the spring pretest measure provides a kind of "treatment vs. control group" comparison. We expect fall students on the posttest to score higher, on average, than spring students on the pretest. A comparison of the scores on the spring pre- and posttests provides a second "control vs. treatment" comparison. We expect scores to be higher, on average, at posttest than pretest. The patched-up design, then, builds in a replication of the treatment effect. Finally, we expect no significant difference between the means of the fall and spring final examinations.

Before presenting the results, selection bias must be considered. Clearly, differences between fall and spring test scores may arise because students enrolling in the two semesters are different. (The data reported below suggest that, at least in terms of final examination scores, students taking the course in different semesters are remarkably similar.) This would directly affect the fall final/spring pretest comparison, but not the spring pretest-to-posttest comparison; the patched-up design provides some protection against possible selection bias. Moreover, exchangeability concentrates attention on the level of achievement and considers the "gain" in achievement a necessary but not sufficient indicator of student outcomes. Indeed, level of achievement is the "bottom line" in an evaluation of exchangeability. Systematic differences between student populations may be acknowledged, but they are not used statistically to equate groups or in any other way



(a) Preferred



(b) Implemented

TRA = traditional

O = observation/measurement

TEL = telecourse

X = treatment

Studenting (S) = student course participation

Fig. 1—Preferred and implemented patched-up designs

to make allowances for differences in knowledge and skills gained at the end of the course.

Within-Site Treatment Effects

We next examine the effect of the telecourse "treatment" in each site. We compare fall posttest scores with spring pretest scores, and spring pretest scores with spring posttest scores, to estimate mean gain. Finally, we compare fall posttest scores with spring posttest scores to determine whether the outcomes in one semester were equivalent to those in the other semester.

Data in Fig. 2 indicate a consistent difference between fall posttest scores and spring pretest scores of about 8 points out of a total possible 50 ($p < .01$). Except for Site E, a similar gain (about 8 points) is found in spring ($p < .01$). Finally, except for Site E, the differences between mean fall and spring posttest scores are negligible and not statistically significant.

At Site E, the mean spring posttest score is 48 out of 50. That is extraordinary. However, the implementation study findings explain the anomaly: the instructor in this course provided students the answers to the midterm and final examinations at the review sessions, and gave "open book" tests.

The evidence, then, is that *The New Literacy* telecourse, as implemented at four sites, significantly increases students' knowledge about computers. Moreover, comparison of fall and spring posttest scores indicates that the course outcomes are remarkably similar in the two semesters across the sites (exception noted). Finally, as might be expected, the design is sensitive to research implementation problems at Site E.

Between-Site Treatment Effects

Although different implementations of *The New Literacy* produced roughly the same mean score differences across sites, a direct comparison of raw mean difference scores is hazardous because the scales of measurement at the different sites were not the same; different tests were used. To compare the consistency (or inconsistency) of mean differences across sites, we put the mean differences on a common metric by dividing the difference by the (pooled) standard deviation. In other words, we converted the mean difference score at each site into a measure of *effect size*. The results of this analysis are shown in Fig. 3.

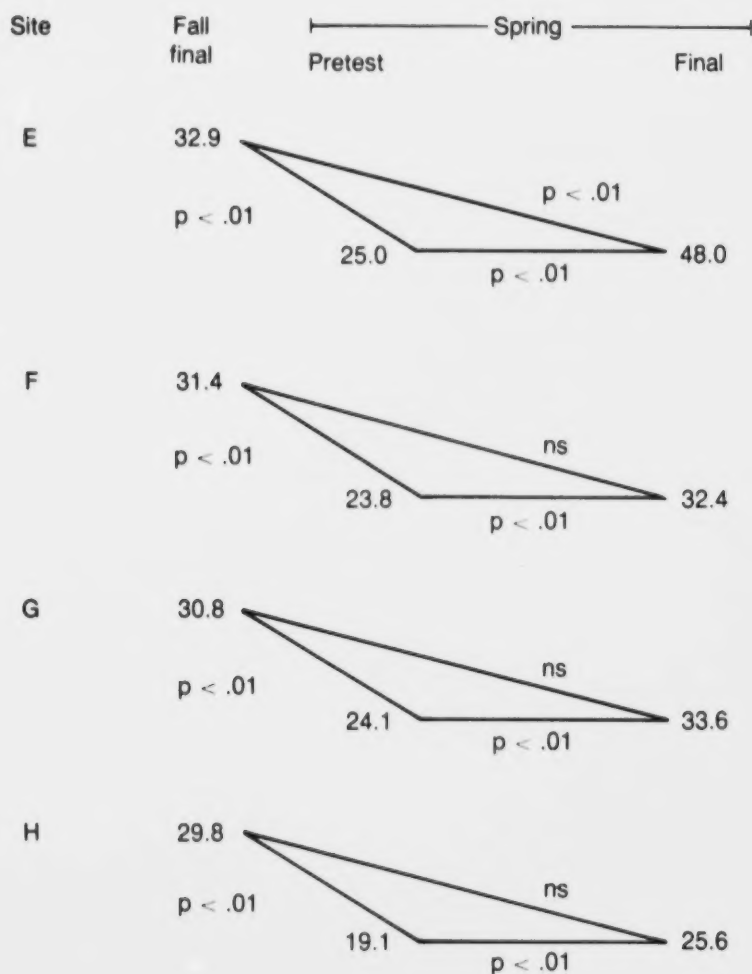


Fig. 2—Within-site mean achievement comparisons

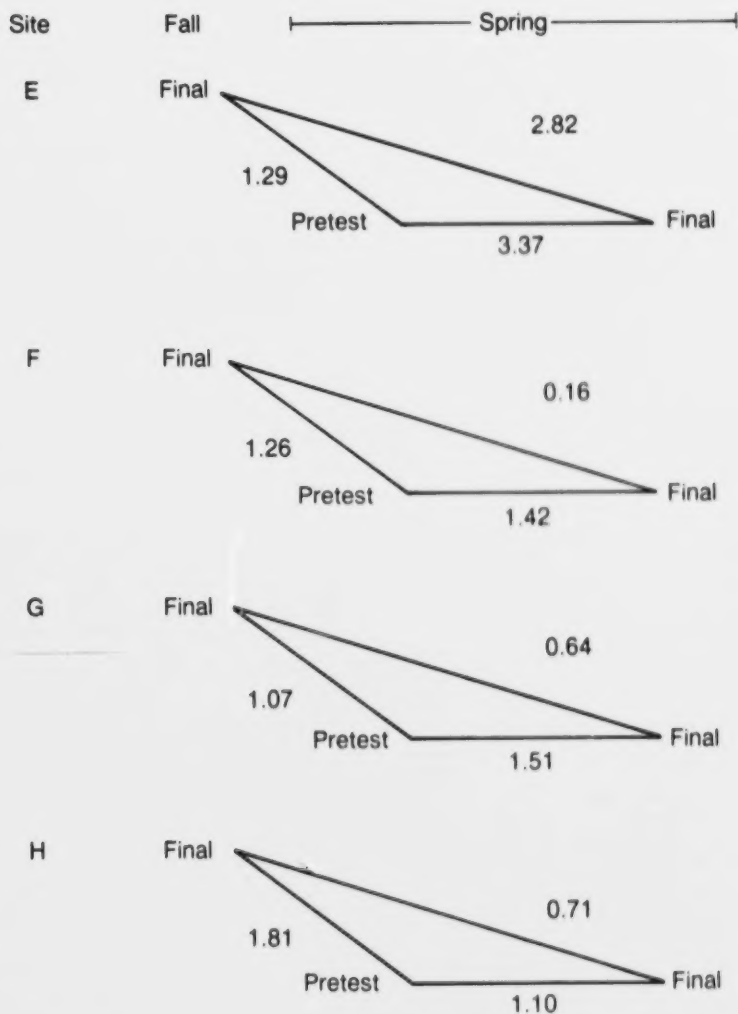


Fig. 3—Between-site achievement effect-size comparisons

Data in Fig. 3 show a consistent effect size of about 1.36 standard deviations comparing fall posttest and spring pretest scores across all sites. Except for Site E, a similar effect size is found comparing spring pre- and posttest scores. Site E, as noted above, is the exception, with an effect size over double that of any other site. Finally, except for Site E, the effect size for the difference between mean fall and spring posttest scores is negligible across the sites. These findings lead us to conclude that, in spite of variation in telecourse implementation within and across sites, the gains in test scores are remarkably stable from one semester to another. More importantly, we find similar effect sizes even though the examinations used at each site were *different*.

If we treat the findings at the sites as replications, we conclude that the magnitude of the mean gain is replicated at each of the sites (Site E's posttest is the exception noted above). This increases our confidence in the patched-up design, and in our interpretation that the telecourse caused the increase, not some outside, extraneous event. In sum, the patched-up design, at least as far as we have tested it, seems to provide reliable, valid, and useful information about the effect of the *The New Literacy* on student knowledge about computers.

To bring additional information to bear on the treatment effect, we compared spring pretest-to-posttest gains using the Rand common items across all telecourse sites (Fig. 4). This permitted us to compare all sites on an *identical* measure, and to add a fifth site, Site D1, to see if the magnitude of the mean difference would replicate at this site as well. The difference between mean pretest and posttest scores was significant at each site ($p < .05$), and the mean gains as measured by effect sizes were consistent across sites, lending further confirmation of the consistency of the magnitude of the treatment effect.

Finally, we looked at correlations and the relative difficulty of unique, locally developed items as compared to Rand common items to determine if the common items that all sites received are typical of the test items administered at each site. Table 9 indicates that Rand items are less difficult at Sites E and H and more difficult at F and G. However, the differences are quite small.

Correlations Between Student Background, Course Participation, and the Rand Common Posttest

Correlations between selected student characteristics (demographic, ability, and education) and the Rand common posttest scores are presented in Table 10. Before interpreting these correlations, a caveat is in order. Correlations based on as small samples as we had at each site have very wide confidence bands. Even those identified as

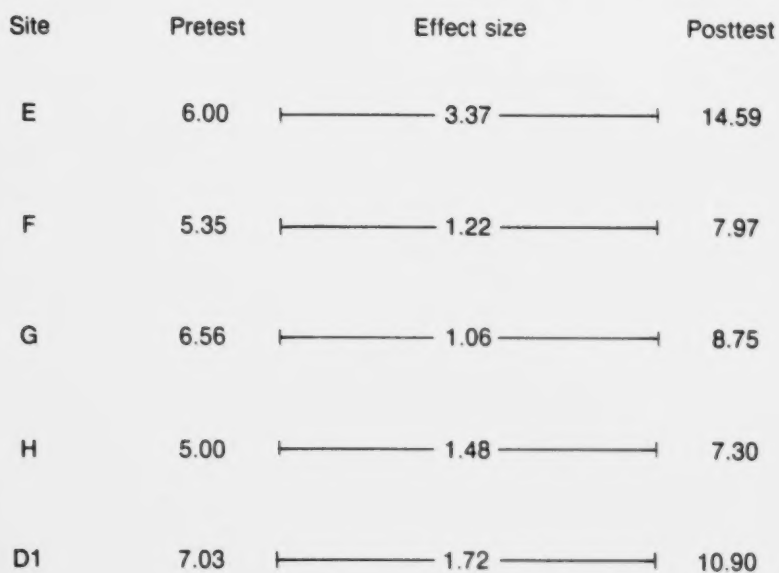


Fig. 4—Spring pretest-posttest comparisons with 15 Rand common items (means and effect sizes)

Table 9
COMPARISON OF ITEM DIFFICULTY BETWEEN
UNIQUE AND COMMON ITEMS

Site	Mean Proportion Correct		p	Correlation
	Unique Items	Common Items		
E	.91	.94	<.008	.65
F	.62	.51	<.001	.62
G	.69	.58	<.001	.53
H	.58	.51	<.111	.36

Table 10
SELECTED CORRELATIONS BETWEEN STUDENT CHARACTERISTICS
AND FINAL TEST SCORES (15 ITEMS)

Characteristic	Site				
	E	F	G	H	D1
	Sample Size				
	29	34	16	20	29
Vocabulary	-.16	.19	.54*	.11	.29
Rand pretest	-.04	.28*	.58*	.34	.00
Grades	.25	.38*(33)	.30	.18	.12
No. of previous computer courses	.08	.29*	.14 (15)	-.50*	.12
No. of previous telecourses	-.51*(28)	.33*	.25	-.46*	-.02
Enroll in regular course if no telecourse	-.01	-.35*	-.70*	-.25	-.11
Work full-time	-.18	-.01	.08	.40*	-.03
Attitude toward telecourses	-.20	-.15	.14	.45*	.21
Hours watching TV per week	.35*(28)	-.03 (33)	.12	-.57*	-.06

NOTE: Exceptions to sample size are in parentheses.

*p < .05.

statistically significant may be spurious. We report the correlations for completeness and interpret them where some degree of consistency is found across sites.

Measures of ability correlate positively with posttest scores. For example, students' reports of their grades are positively, although rarely significantly, correlated with posttest performance.

A curious pattern of correlation emerges for telecourse-related variables. Students willing to take *The New Literacy* as a traditional course if the telecourse were not offered tend to score lower on the posttest than students not willing to make the shift. Finally, students with prior telecourse experience tend to score lower on the posttest than students without such experience, although the correlation pattern is mixed. We have no explanation for these two findings, especially if they are taken together.

Correlations between student course participation ("studenting") variables and Rand common posttest scores are presented in Table 11 with the same caveat (wide confidence bands) mentioned above. Three studenting variables are consistently related to test performance across the five sites: (a) percentage of textbook reading assignments completed, (b) number of class meetings attended, and (c) rating of instructor's preparation. Taken together, these findings highlight the importance of text materials in the telecourses, and the benefit of class meetings to review the material to be covered on the test. These interpretations are consistent with the implementation study finding that instructors emphasize text material and use class meetings to review for tests.

Attitudes Toward Telecourses

Spring pretest-to-posttest changes in attitude toward telecourses across the five sites are presented in Fig. 5. While students, on average, held positive attitudes toward telecourses across the five sites, these positive attitudes consistently declined slightly (except for Site H). Nevertheless, we interpret the findings to show that students hold positive attitudes toward telecourses both before and after the course. Perhaps the changes reflect the imperfections in telecourses and their local implementation, or perhaps they reflect regression effects.

Table 11
SELECTED CORRELATIONS BETWEEN STUDENT COURSE PARTICIPATION AND FINAL TEST SCORES (15 ITEMS)

Participation	Site				
	E*	F	G	H	D1
	Sample Size				
	63	43	22	16	57
Record programs	.03	.32 ^b	-.23	.28	-.05
View programs > 1 time	-.26 ^b (62)	.04	.27	.11	.23 ^b
Rigor of telecourse vs. traditional course	-.03 (59)	-.27 ^b (42)	-.20	.32	-.01
Hours study per week	.20 (60)	-.05 (42)	-.25	.22	.02
No. of meetings attended	.42 ^b (60)	.00	.30 (21)	.44 ^b (15)	.29 ^b
Rating of helpfulness of class meetings	.51 ^b (62)	-.42 (11)	-.07	.28 (15)	.17 (56)
Instructor preparation	.35 ^b	.12	-.04 (21)	.33	.25 ^b
Take notes while viewing	.06	.00 (42)	-.40 ^b	-.14	-.01
Understand videolesson	.05 (62)	.09 (11)	-.42 ^b	.52 ^b (15)	-.13 (55)
Percent assignments read	.08	.11	.42 ^b	.31	.41 ^b
Take notes while reading	.00	-.02 (41)	-.43 ^b	.13	.26 ^b
Understand reading	-.01 (62)	-.09 (11)	.10	.48 ^b	-.07 (56)
Answer questions in reading	-.03	.06	-.09	.37	.30 ^b

NOTE: Exceptions to sample size are in parentheses.

*Instructor provided answers to test questions; very little variance in posttest due to ceiling effect.

^bp < .05.

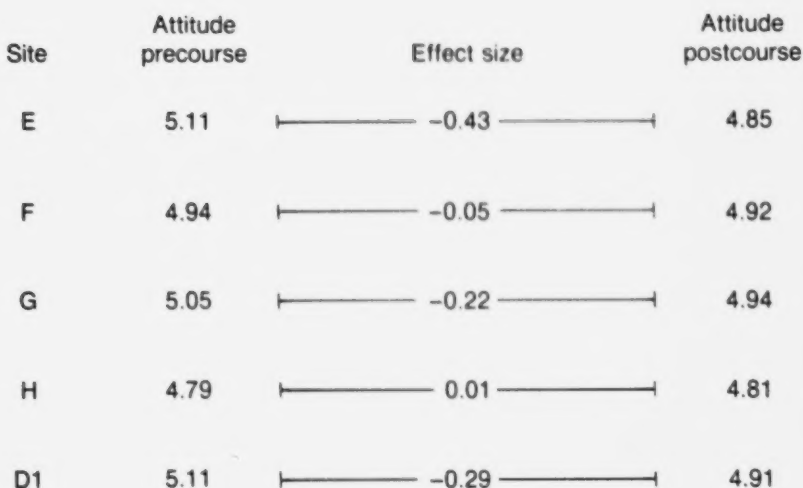


Fig. 5—Spring attitude precourse and postcourse comparisons
(means and effect sizes)

CONCLUDING COMMENTS

The evidence presented here leads us to conclude that, methodologically, the patched-up design provides a fairly unobtrusive, cost-effective, and time-minimizing design that produces reliable, valid, and useful information for evaluating student outcomes from telecourses. We speculate that by combining a traditional course and a telecourse in the patched-up design, in some approximation to Fig. 1(a), this design will provide relatively unobtrusive, cost-effective, and time-minimizing data on the *exchangeability* of student outcomes. A caveat, however, is in order. We speak of "relatively" unobtrusive because the design requires pretesting, something not normally done in college classes. Further, no evaluation is unobtrusive. As the implementation study (Sec. III) showed, college instructors are especially sensitive to even the least intrusive evaluation design. Without adequate orientation and continued support, they may create, knowingly or unknowingly, severe problems for the conduct and interpretation of evaluation.

V. RECOMMENDATIONS AND CAVEATS

As we have indicated, the central question that an evaluation of student telecourse outcomes must address is: "Are the knowledge, skills, and attitudes acquired by students in a telecourse *exchangeable* for the knowledge, skills, and attitudes acquired by students in an equivalent traditional course?" In Sec. II, we noted what exchangeability implied for evaluation design. First, the design should be *comparative*; the purpose of the traditional course is to provide a standard against which to evaluate telecourse outcomes. The "treatment" and "control" groups in an evaluation—the telecourse and traditional course—are each defined to include both the *curriculum* and its *local implementation*. Second, student characteristics and self-selection cannot be disentangled from local implementation. An evaluation should yield an estimate of student gain from the telecourse, but the magnitude of this gain should not be indexed to the magnitude of gain in traditional courses because it is not appropriate to equate the two groups of students. Third, an evaluation should use designs that balance internal validity (correct attribution of causes for observed tele-/traditional course differences) and external validity (inference to the larger set of telecourses and traditional courses not observed in the study).

In this study of the feasibility of evaluating student outcomes from telecourses, we identified several appropriate experimental and quasi-experimental designs for studying the exchangeability of student outcomes in telecourses for outcomes in traditional courses. However, research implementation problems (not methodological limitations) hampered our feasibility study from the beginning. Neither of the designs we implemented at *The Constitution* sites (the pretest-posttest design and the posttest-only design) provides direct evidence on the exchangeability of student outcomes from tele- and traditional courses. For *The New Literacy*, Site D offered equivalent tele- and traditional courses, but the dropout rate was so high in the traditional course that our findings were less useful than we had hoped. Finally, only the "patched-up" design proved workable, and only at the other four *New Literacy* sites.

The many problems we encountered during this feasibility study enable us to offer guidance to researchers who plan to evaluate student outcomes of telecourses more systematically and comprehensively:

- Oversample . . . to excess.
- Obtain ample lead time.
- Pretest comprehensively.
- Study course implementation in depth.
- Do not underestimate the intrusiveness of evaluation research.

In proposing these five "commandments," we do not pretend to godly wisdom. Rather, we are highlighting sources of devilish difficulty in conducting evaluation research in the real world of "distance teaching."

1. OVERSAMPLE . . . TO EXCESS

To obtain sufficient data to evaluate adequately the exchangeability of student outcomes from tele- and traditional courses, it is essential to start with far more sites than might appear necessary on scientific grounds. Our field experience amply demonstrated the two main reasons: first, much can happen before the study ever gets under way that may lead to the loss of carefully chosen sites; and second, the evaluator's ability to regulate how administrators and instructors implement the tele- and traditional courses is, in reality, quite limited—the structure of American higher education, and of "distance teaching" especially, is so informal that it is difficult to predict the degree to which courses will actually operate as designed.

Previous examples amply validate these general assertions. The troublesome question they raise is whether the requirement to oversample will render evaluation unfeasible. Certainly it will increase the cost of evaluation and make it unwieldy. A more basic problem, though, is whether a sufficiently large sample of appropriate sites exists in practice.

In this study, we would have had no difficulty expanding our sample of *New Literacy* sites for inclusion in the patched-up design; as we implemented it, only telecourses, not traditional courses, participated. But it would have required far greater time and effort to locate additional campuses to find tele- and traditional course comparisons for the non-equivalent control group design. With regard to *The Constitution*, it is not clear that a sufficient sample of telecourse treatments exists on which to conduct an evaluation. No matter how much time is devoted to recruitment, the fact is that, so far, very few campuses throughout the country appear to be using the course. And the limited number of *Constitution* sites, of course, makes it far more difficult to include this course in designs that compare outcomes from tele- and traditional courses.

Thus, the requirement of liberal oversampling presents an evaluation dilemma that time and money will not necessarily resolve. Unless implemented in artificial conditions, certain telecourses—not just specific courses but entire *genres* of courses—may be extremely hard to find in adequate numbers to make an evaluation worthwhile.

2. OBTAIN AMPLE LEAD TIME

The study lacked adequate lead time (approximately two weeks to recruit and orient sites, and one month to develop instrumentation). Without doubt, the study suffered accordingly. For example, several desirable *Constitution* sites began spring classes before we were able to complete instrumentation, and so could not be included. In addition, we were unable to contact potential sites or to locate administrators or instructors during the Christmas break. In one instance, as described earlier, our inability to contact the instructor (who was on vacation) led to our including a site (H) entirely on the basis of incorrect information supplied by the telecourse administrator. As a result, the value of the student outcome data we eventually obtained was diminished. Coordination with several other potentially valuable sites simply could not be finalized in the short lead time available.

The value of sufficient lead time extends well beyond the obvious desire to incorporate all appropriate sites and to avoid embarrassing snafus. It is not enough merely to arrange for a site to participate in an evaluation study. Participants must be informed and educated to the study's goals and methods, with full recognition that what the evaluator requires may well be alien to their customary routines. Consider the matter of examinations. Instructors in higher educational institutions are generally free to devise and modify their own tests without interference or even oversight from anyone (or, at least, certainly no one beyond their department chairman). Our study challenged this academic custom. At several sites the instructors were obviously upset by our role in test design. Especially in those telecourses where the instructor generally did little more than prepare students for and administer exams, an evaluation intended to be unintrusive appeared quite intrusive.

To obviate this difficulty—which is to a large degree inherent in the requirements of evaluation work—we believe that evaluators should begin working closely with instructors at least three months before the course begins. If they are to participate in a design with its own instrumentation, the instructors should either contribute directly to, or at least be made comfortable and knowledgeable about, all achievement

tests (we will comment separately below on other instruments). If an instructor is intransigent and will not approve an examination created by an outside evaluator, he or she should be encouraged to help the evaluator develop the tests necessary to include the site in a patched-up design. Either way, the instructor must be consciously cultivated to become a willing participant and to be ready, also, to defend the tests. Otherwise, as our experience demonstrated, an alienated instructor may seek to undermine, scapegoat, or seriously complicate administration of the tests. The hurried manner in which—because of short lead time—we were compelled to review examinations with instructors (especially in *The New Literacy*) worked to no one's advantage, bred antagonisms, and, ultimately, compromised the integrity of student outcome data.

3. PRETEST COMPREHENSIVELY

Because pretests are critical to evaluation of student outcomes, their administration is of central concern. Pretest administration presents special difficulties in "distance teaching." These must be brought to the telecourse administrator's and instructor's attention long before the course begins.

Because on-campus attendance requirements in telecourses are usually minimal, most administrators encourage students to attend a preliminary orientation session during which they meet the instructor, learn about course requirements and methods, and, sometimes, hear general advice on the special demands of studenting in telecourses. In actuality, orientation attendance is very loose; students can usually appear at other times to pick up essential syllabi or handouts (if they have not already received them by mail) and to buy books. They need not attend class until the midterm review session or even the midterm examination itself, if they so choose. These loosely structured attendance requirements facilitate late registration by students who do not become aware of a desired course until it is under way.

If obtaining students' attendance at the first class meeting is a problem, administering a pretest to those who do attend also raises difficulties. Most telecourse instructors, it would appear, do not normally administer a pretest of their own. They use the orientation session (usually an hour or so) rather casually to introduce themselves, the course requirements, explain why the subject matter is important, and either affirm the advantages of, or reveal their own ambivalence about, telecourses in general. Hence, administration of a 45-minute pretest can dramatically change customary routines (such as extend the

orientation an extra 45 minutes, or limit the instructor to a 15-minute introduction of himself and the course). These changes can jar student as well as instructor expectations if—as was sometimes true in the study because of insufficient lead time—students had not been warned in advance to expect a pretest (or, indeed, to expect outside involvement at all).

Although pretest administration went relatively smoothly at several sites (even these reported minor student dissatisfaction), at other sites instructors, students, and administrators complained about the extent of intrusion. At one site (E), as described earlier, the administration of the pretest was a fiasco. The administrator's surprise decision to physically separate pretest administration from telecourse orientation was a logistical error that significantly reduced the pretest sample.

Because of the problems and uncertainties in pretest administration, *we did not have pretest scores for approximately half of the students who finished the telecourses at all sites.* This issue obviously needs to be addressed and can be done so in several ways, all of which presuppose substantial lead time.

Even if administrators cannot, for financial reasons, rigidly require attendance at orientation sessions (for fear of losing clientele), they can require students to take a pretest as a course requirement. To do so, they must either lengthen the initial orientation session (as some sites did), shorten the amount of time in which instructors formally talk to students about course objectives (some of this can be dealt with by mail), or provide for some other arrangement consistent with institutional policy. Equally important, administrators must arrange to pretest—on-campus and under suitable direction—late registrants who often form a substantial portion of final enrollees. This will not be easy for hurried administrators and staff, but comprehensive pretesting is simply essential.

To make the pretest truly mandatory, it may be necessary to pay students to take it, or to maintain a member of the evaluation team on campus for as much as several weeks to ensure proper administration. In our view, however, the real key to effective pretesting is to devote sufficient *early* effort to informing administrators and, especially, instructors about the purpose of all instruments, the necessity of maintaining their integrity, and the desirability of evaluation per se as integral to their pedagogical effort. If the pretest is perceived as an alien, externally imposed requirement rather than as integral to the teaching/learning process, not only are the data themselves likely to be jeopardized during test administration, but relations between evaluators and instructors will get off to a bad start that may seriously compromise the research design.

4. STUDY COURSE IMPLEMENTATION IN DEPTH

The study focused on design and student outcomes, but implementation of the course was considered an integral part. Indeed, implementation issues intruded so often as to make them central to interpretation of outcomes. Future telecourse evaluation should place special attention on implementation, not only because it influences student outcomes but also because it illuminates a variety of important general issues concerning the future of "distance teaching" in American higher education. This requires on-site monitoring of all aspects of telecourse implementation.

Instructor Variability

The most basic implementation subject, at least for studies of student outcomes, is *how* the course was actually taught. The way a treatment is defined is not necessarily the way it is delivered. At first glance, it may appear that the structure and content of telecourses are relatively simple and fixed compared with traditional courses. Telecourses are designed so that students can meet on-campus infrequently; they feature textbooks designed specifically to accompany the videolessons. However, instructors can introduce tremendous variation into seemingly standardized course formats. One *New Literacy* instructor met with students on campus four times for a total of seven hours, whereas another met eighteen times for fifty-four hours. Several taught the BASIC computer language and several did not. One *Constitution* instructor administered four examinations whereas the others gave only two. Two instructors graded students entirely on the basis of test performance, whereas the other counted tests for only 40 percent of the students' grades. In short, just because telecourses seem more like predetermined "packages" than traditional courses does not mean that instructors will treat them that way.¹ There is much room for variability in course formats and procedures, and these variations may influence how and what students learn.

Centrality of Videolessons

A particularly important implementation issue is how central the videolessons are to the instructor's pedagogy (we beg the more difficult

¹A problem that is perhaps inherent in the production of sophisticated telecourses is a tendency for their creators to view them as self-contained, self-executing, and, to a large extent, teacher-proof curricula. The realities of implementation—of how these products will be used in real-world education settings—seem relatively mundane and secondary to their purposes.

question of how central the videolessons are to what students actually learn). Do the instructors integrate the videolessons into their in-class teaching? Do they tell students that the videolessons are important to watch, advise them on how to learn from the videolessons, or hold the students in any way responsible for learning what the videolessons teach? In short, do the videolessons matter at all in the instructor's implementation of the telecourse?

In four of the five telecourse sites for *The New Literacy*, the instructors did little more than advise students at the beginning of the semester to watch the videolessons, and did not refer to them again. The textbook drove the course, supplemented to a far lesser extent by the telecourse study guide. Only the telecourse instructor at Site H, who met his class weekly, referred at all regularly to the videolessons. And he did so far less after the first few weeks when, in his judgment (and, apparently, that of his students), the videolessons became repetitive and boring.

The videolessons and accompanying textbook and study guide in *The Constitution* were of such a different nature from those in *The New Literacy* that true comparison of the two telecourses is impossible. The textbook for *The Constitution* (Friendly, Elliott) was written in a chatty, popular style and overlapped only intermittently with the videolessons. The study guide (McKenna), on the other hand, was very much a textbook: larger, more comprehensive, more scholarly, more insistently instructive than the nominal textbook. In *The Constitution*, as in *The New Literacy*, the "text" (i.e., McKenna) drove the substantive content of the course.

There were, however, two main differences in the way that the videolessons were used in *The Constitution* versus *The New Literacy*. First, in two of the three *Constitution* sites, instructors sought to use in-class meetings to stir classroom discussion along lines similar to the style of the debates featured in the videolessons. Second, at one site (A), students' ability to emulate the style of the videolessons in classroom debate and in term papers figured centrally in determining their grades. Thus, videolessons were more integral in the conduct of *The Constitution* than of *The New Literacy*. Nonetheless, at all sites written materials overwhelmingly supplied the body of knowledge that students were expected to master.

How to motivate faculty to make more active use of the videolessons is problematic for curriculum developers and evaluators. Some instructors appear highly ambivalent about telecourses as "equivalent" bodies of instruction; they may fear reduction of the teacher's role. Faculty may require elementary instruction on how to integrate videolessons into student learning and into the development of evaluative instru-

ments. One thing seems certain: if faculty do not systematically tie the videolessons to the written materials and tests, it is hard for anyone to regard telecourses as anything more than traditional, old-time correspondence courses.

In the end, a lingering ambiguity surrounds the concept of the "telecourse." The term inevitably implies that the videolesson carries the instructional burden of the course. It may be that the term is a misnomer: we may need a new label for describing what are, in reality, complex, highly diversified, multimedia educational packages in which television plays an as yet undetermined pedagogical role.

Purpose of Review Sessions

An equally important component of course implementation is the examination review session which, at most sites, is a separate class session convened solely for that purpose. There may be an inherent problem in these review sessions: they seem to lend themselves to a spoonfeeding type of test preparation fundamentally different from that in most traditional courses. Following the orientation, the students generally do not see the instructor until the week before, or even the day of the midterm examination. The instructor does not know the students personally or have experience in conversing with them. And few instructors anywhere are likely to be skilled or confident about concisely outlining an entire half semester's work in one class session. The instructor's most natural response in this unfamiliar situation—aside from simply entertaining students' questions—may be to review in great detail what is to appear on the test, highlighting precise textbook pages and even paragraphs from which test questions may be drawn. At least in that way he knows he will be helpful, whereas it is otherwise unclear what he should do. It may also be the case, especially with part-time instructors, that there are considerable enrollment pressures to make students happy so that they will return to take more telecourses, thereby guarding the faculty member's part-time slot. These pressures may further redirect the midterm and final examination review sessions in the direction of spoonfeeding.

All of this is to say that we dare not consider our most extreme example of spoonfeeding—at Site E, where the instructor actually gave the students the exam answers in advance—to be entirely atypical. We may be dealing, rather, with a continuum in which Site E is at one end and several others are not too far away. And the problem may be inherent in the most common format of telecourse implementation.

Political Influences

There remain a number of lesser order operational concerns,² but we turn now to wholly different implementation issues, those dealing with what may loosely be termed "political" questions. They involve decisionmaking based, at least in part, on considerations of power and ideology. Political decisionmaking was evident in a variety of ways: in one site's decision to require telecourses to meet every week, just like traditional classes; in the varying degrees of control and veto power exercised by departments over which telecourses to offer and how to staff them; in the designation of special course labels to describe or differentiate telecourses from traditional courses; in the addition of telecourse requirements beyond the A/CPB curriculum package; in the granting of fewer course credits to telecourses compared with traditional courses; in the roles played by administrators at several levels in the campus hierarchy in encouraging or discouraging telecourses on campus; and so on.

In short, the telecourse is a controversial, emotionally charged issue in American higher education. We should not underestimate the threat it represents—and the more sophisticated the telecourse, the greater the threat.³ Therefore, we should expect to find no prototypical telecourse and, moreover, we should expect to experience difficulty in gaining cooperation from departments in comparing a telecourse with a traditional course. Politics, consequently, can undermine an evaluation of exchangeability.

Unfortunately, the study barely skimmed the surface of this fascinating topic. We cannot determine with precision from our data how political issues that are *generic* to telecourses, *specific* to *individual telecourses*, and *specific* to *individual campuses* affected implementation. Not to consider these matters in future evaluations would be naive.

5. DO NOT UNDERESTIMATE THE INTRUSIVENESS OF EVALUATION RESEARCH

Evaluators naturally have a high opinion of their objectives and methods; they view their requests for change as minimal and obvious to

²For example, does the site require its own posttest questionnaire? If so, does it overlap with the evaluator's, and can students reasonably be asked to fill out both in class?

³American higher education institutions have traditionally prided themselves on their local and regional character and responsiveness. Telecourses, by contrast, represent a nationalizing, translocal and transregional influence. Hence, the relation between telecourse producers and educators is inevitably tension-ridden. Even if telecourses do carry their own weight financially on campus, they challenge both the local initiative ethos and survival needs of America's higher education industry.

anyone who even marginally understands scientific method (as, presumably, any college professor ought). What evaluators may be insufficiently sensitive to, however, is the strongly engrained tradition of autonomy in American higher education. The study demonstrated just how laissez-faire the system is, particularly in the classroom, where generally no one but the instructor exercises much control or even knows what transpires. Time and again we confronted administrators who truly had no understanding of what took place in the classroom, or who did not discover flagrant abuses until well after the fact.

Under these conditions, it would certainly come as no surprise if the presence of an evaluator (especially an outside evaluator) appeared intrusive. This would hold particularly if the instructor had not been informed about the evaluation before he agreed to teach the course. In such cases it is not surprising that looking over their shoulders and asserting control over testing may well be disquieting, threatening, and alienating. Moreover, our various tests (which tended to be more difficult than some instructors were accustomed to) held real potential for intimidating students, even to the point of their dropping the course.

Nor should one understate the threat which the telecourse may represent to the full-time college instructor. The threat to his security has been present since the days of "talking head" educational television—why would students want to hear a mediocre lecturer if they could, conceivably, have the very best lecturers on videotape? The threat in the post-"talking head" era is somewhat different—"lecture" is no longer the preferred style of presentation—but it is no less real. Not only are A/CPB-type courses far more attractive and familiar in style to students who have grown up with TV, they subordinate the teacher to the telecourse curriculum package. There is, of course, no reason why telecourses must be seen as displacing or downgrading the instructor's role. They can be viewed as fostering new opportunities for transforming the teacher's prime role, especially in colleges which rely heavily on large lecture classes. There is little indication, however, that teachers want to have their classroom roles radically transformed, especially by the advent of impersonal new instructional technologies. And the fact that telecourses have been conceived so far primarily to enhance "distance teaching" may only increase the fear of regular instructors that non-educator, media sophisticates are out to steal their clientele with a bogus product.

In sum, it would be naive not to recognize that an evaluation *can* cause a telecourse instructor considerable anxiety, and that an evaluation *may* further alienate an already alienated teaching environment. This conclusion reinforces the need for evaluators to make early contact with potential sites, and to discuss all components of the

evaluation with instructors and administrators in order to secure their knowledgeable cooperation. It may be advisable to invest substantial resources in paying instructors to participate in seminars sponsored by the evaluation team to make them feel comfortable and, equally important, to give formal recognition and status to the key role they will play. And it may also be advisable to supplement instructors' salaries for their additional time and effort in accommodating the evaluation team during the conduct of research (contingent, of course, on the execution of agreed-upon procedures).

The other group which may perceive an evaluation as intrusive is, of course, the students. Some difficult ethical issues arise when students who pay for a course are asked, or compelled, to fulfill requirements that are tangential to that course, especially if they were not informed about them before they registered. And if students resent the "extra" demands made upon them to the point of becoming antagonistic to the instructor or administrator, and perhaps even dropping the course, practical considerations arise that may jeopardize the evaluation: students may not perform as well as they otherwise might, or the evaluators may be asked to leave. These potential problems never reached the crisis point in the study, but at several sites there was enough student disquiet to suggest that they could become troublesome in future evaluations. It may therefore be advisable, and ethically appropriate, for the evaluator to pay part or all of students' fees and book expenses in order to secure their willing cooperation. It is certainly essential to inform students about the evaluation before they register for the course.

The ideal solution to this potential problem, however, involves not money but a change in institutional procedures. Higher education administrators who offer as new and controversial an experiment as telecourses need to be persuaded to build in sophisticated data collection as part of their institutional routines. If they can be helped to implement testing procedures at minimal cost, the institution would have in-house a potent tool to pursue self-improvement, and this potential source of student alienation would significantly diminish.

CLOSING OBSERVATIONS

Telecourse evaluation is an intellectually hazardous enterprise. The structures and procedures common in higher education, especially in "distance teaching," do not lend themselves to easy regulation, manipulation, or control.

Despite abundant difficulties, we believe it is possible to monitor implementation so as to secure reliable data on student outcomes without intruding unduly into normal pedagogical routines. Certain conditions, however, must be observed:

1. A sufficient number of sites must be available to oversample each course.
2. Extensive lead time must precede implementation of the research so that site administrators and instructors can be closely involved in design and instrumentation.
3. Pretesting must be mandated and carried out for all students.
4. Implementation data must be integral to analyses of student characteristics. Researchers should be on-site to gather information on four key matters:
 - How was the course taught (format and requirements)?
 - Did the instructor integrate the videolessons into his teaching and evaluation?
 - How were the review sessions conducted?
 - What political factors affected course implementation?
5. Via education and, perhaps, monetary inducements, instructor and student willingness to adhere to all evaluation procedures must be secured.

In sum, serious impediments to implementing rigorous evaluations of telecourses are many, but they are not insurmountable.

Appendix

RAND COMMON ITEMS

The 15 common items used in the patched-up design were selected from a 200-item objective test developed specifically for this course by the Rand research team. To create these items, we first developed a course outline that delineated the important material in the course. A team of two researchers independently watched the 26 half-hour shows and listed the major concepts and facts covered in them. These independent concept lists were merged to create a concept outline for each show. Next, we divided the shows into two groups (approximately 13 shows in each), and the researchers read the textbook chapters corresponding to their assigned shows. Thus, concept outlines were supplemented with material from the textbook. The resulting course outline distinguished between information presented in the video alone, in the textbook alone, or in both video and text. Since the shows, in our opinion, were not equally important, we further distinguished 14 shows as important and 12 shows as less important. For each show, we prioritized the covered concepts and facts according to their importance in the chapter.

In developing the questions, we aimed at a balance between questions that tested facts (factual) and those that required students to draw inferences, transfer to new material, and solve problems (conceptual). We also attempted to cover material presented in both the text and video, and in the text or video alone.

With these goals and our content outline as a guide, we examined a bank of test items created for *The New Literacy* telecourse by the Southern California Consortium for Community Colleges (with assistance from Interwest Applied Research).¹ With this item bank as a starting point, we extensively modified items appropriate for the content we wanted to test and wrote new items. Careful attention was paid to an item's psychometric properties (e.g., short, concise stems, appropriate distractors).

¹Questions are in multiple-choice format and are classified as belonging to one of five different skill levels—recall of fact; recall of implications or interpretations; application of facts, implications, or interpretations; inferential reasoning; and synthesis and evaluation of information. Questions are based on the lesson objectives as they are specified in both the video segment and text assignment. Users of *The New Literacy* could copy these items for testing student achievement. At least three of our study sites used this item bank to construct student exams.

The resulting content by process (fact or conceptual) by media (text, video, both) grid of test items had roughly equal numbers of factual and conceptual items. We were not able to achieve a balance between items from the text, video, or both because most of the important concepts and facts were covered either in the text alone or in both the text and video. Few important concepts were presented in the video alone.

Five of the 15 common items are listed below.

1. If an application program is allocated insufficient real storage on a computer with virtual memory, the most likely result is:
 - (1) The number of page-in operations will drop dramatically.
 - (2) The system will revert to segmentation techniques.
 - (3) Program execution speed will drop to unacceptable levels.
 - (4) Important data will be lost because of read errors.
2. Which statement is true?
 - (1) Computers can only carry out those activities they have been programmed to do.
 - (2) Computers designed with adequate memory can run without programs.
 - (3) Users of application software must know how to program in the language used to write the application.
 - (4) Advanced computers can write their own programs without programmer assistance.
3. Which of the following is LEAST supported by evidence?
 - (1) Computerized inventory control procedures save manufacturers large amounts of money.
 - (2) Small computers substantially improve the achievement of high school students.
 - (3) Computers provide easy access to law enforcement records.
 - (4) Computers substantially improve the engineering of automobiles.
4. What computer application illustrates transaction processing?
 - (1) Producing monthly telephone bills.
 - (2) Producing state unemployment checks.
 - (3) Scoring students on standardized achievement tests.
 - (4) Typing reports on a word processing program.

5. Which statement is most accurate about 8-bit and 32-bit processors?
- (1) The 8-bit processor requires more operations to complete the same task than the 32-bit processor.
 - (2) Only 32-bit (or greater) processors can add more than two numbers in a single operation.
 - (3) The 8- and 32-bit processors perform similar internal operations, but the 32-bit processor calculates 64 times faster.
 - (4) The 32-bit processor can do many operations that the 8-bit processor can't.

REFERENCES

- Burstein, L. "Issues in the Aggregation of Data," *Review of Research in Education*, 1980, Vol. 8., pp. 158-236.
- Campbell, D. T., and J. C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," in N. L. Gage (ed.), *Handbook of Research on Teaching*, Rand McNally, Chicago, 1963.
- Clark, R. E., and G. Salomon, "Media in Teaching," in M. C. Wittrock (ed.), *Handbook of Research on Teaching*, Third Edition, Macmillan, New York, in press.
- Cook, T. D., and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand McNally College Publishing Co., Chicago, 1979.
- French, J. W., R. Ekstrom, and L. Price, *Kit of Reference Tests for Cognitive Factors*, Educational Testing Service, Princeton, New Jersey, 1963.
- Schramm, W., *Big Media, Little Media: Tools and Technologies for Instruction*, Sage Publications, Beverly Hills, California, 1977.
- Sherwood, C. C., J. N. Morris, and S. Sherwood, "A Multivariate, Non-randomized Matching Technique for Studying the Impact of Social Interventions," in E. L. Streuning and M. Guttentag (eds.), *Handbook of Evaluation Research*, Sage Publications, Beverly Hills, California, 1975.
- Streuning, E. L., and M. Guttentag (eds.), *Handbook of Evaluation Research*, Sage Publications, Beverly Hills, California, 1975.
- Yin, R. K., "The Case Study as a Serious Research Strategy," *Knowledge: Creation, Diffusion, Utilization*, Vol. 3, 1981, pp. 41-49.

RAND/R-3422-CPB